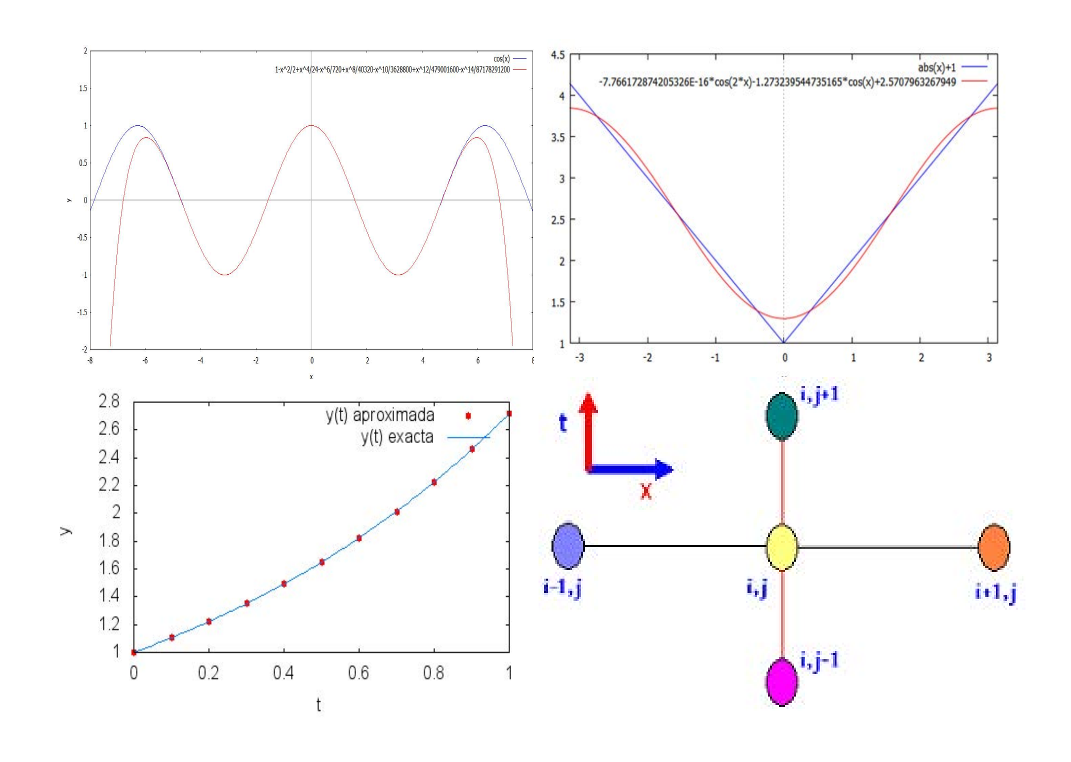


CÁLCULO NUMÉRICO
Teoría, problemas y algunos programas con Maxima



Antonio Vigueras Campuzano
Departamento de Matemática Aplicada y Estadística
Escuela Técnica Superior de Ingeniería Industrial
UNIVERSIDAD POLITÉCNICA DE CARTAGENA





Cálculo Numérico: Teoría, problemas y algunos programas con Maxima

Antonio Vigueras Campuzano





CRAI Biblioteca
Plaza del Hospital, 1
30202 Cartagena
968325908
ediciones@upct.es

Primera edición, 2016

ISBN: 978-84-608-7867-4

Imagen de la cubierta: Representaciones gráficas con Maxima.



Índice general

Prólogo

La asignatura Cálculo Numérico se plantea como una introdución al estudio de los métodos numéricos. En el caso del Grado en Ingeniería en Tecnologías Industriales de la Universidad Politécnica de Cartagena, se estudia en el primer cuatrimestre del tercer curso. Se trata de una asignatura de 6 ECTS, cuyo objetivo es que el alumno conozca y sepa aplicar los métodos numéricos básicos en situaciones concretas propias de su titulación, así como capacitarle para que pueda preparar y manejar algoritmos y programas de cálculo para la resolución de problemas prácticos, a la vez que comprenda las posibilidades y limitaciones de las técnicas numéricas utilizadas.

Como es sabido, las leyes que rigen buena parte de los problemas estudiados en esta titulación, como transmisión de calor, dinámica de sistemas mecánicos, planificación y control de trayectorias de un robot, análisis de circuitos eléctricos, dinámica de fluidos y medios continuos, deformaciones de sólidos, propagación de ondas, cálculo de estructuras, etc., se modelizan matemáticamente y a menudo acaban traduciéndose en sistemas de ecuaciones lineales o no, problemas de valor inicial o de contorno para ecuaciones diferenciales ordinarias o en derivadas parciales, por lo que no cabe duda de que el conocimiento de estos tópicos, al menos en sus aspectos más básicos y elementales, resulta imprescindible en la formación del ingeniero. Ahora bien, como es sabido pocos son los problemas señalados que pueden abordarse únicamente con las herramientas matemáticas vistas hasta este momento, por ello se hace necesario introducir unos conocimiento mínimos de Cálculo Numérico. En palabras de Henrici, podríamos definir el Cálculo Numérico como la teoría de los métodos constructivos en Análisis Matemático. Se hace especial énfasis en la palabra constructivo ya que, propuesto un problema matemático, además de estudiar la existencia de solución, hay que dar un procedimiento para calcularla de modo explícito. Esta solución se construirá mediante algoritmos, entendiendo por tal una especificación no ambigua de operaciones aritméticas en un orden prefijado. Dos serán los objetivos esenciales de esta disciplina: 1) Dado un problema matemático, encontrar algoritmos, a veces llamados métodos numéricos, que bajo ciertas condiciones permitan obtener

una solución aproximada del problema propuesto; y 2) Analizar las condiciones bajo las cuales la solución del algoritmo es próxima a la verdadera, estimando los errores en la solución aproximada. Se trata pues de encontrar métodos aproximados para resolver todo tipo de problemas matemáticos y analizar los errores producidos en estas aproximaciones, si bien dada la amplitud de la materia ha sido necesario elegir determinados contenidos en detrimento de otros. En esta elección se han tenido en cuenta, fundamentalmente, las necesidades de los estudiantes de la titulación, intentando que los conceptos estudiados en esta asignatura sean de utilidad y sirvan para poder aprender otros por su cuenta cuando sea preciso. Pero, al mismo tiempo, se busca ofrecer un curso coherente y estructurado, para que el estudiante no perciba el estudio de esta signatura como una mera colección de técnicas y recetas para resolver problemas, sino que sea también consciente del significado de los diferentes métodos y conozca sus limitaciones y ámbitos de aplicación, para que sea capaz de decidir cuando un procedimiento es adecuado o no lo es. En resumidas cuentas, aunque un ingeniero no es un matemático, por lo que no tiene obligación de conocer el significado profundo de la materia, en particular los detalles más técnicos y sutiles, sí es un usuario avanzado, especialmente aquellos que desarrollarán su actividad profesional en el campo emergente de la I+D+i, tanto en instituciones públicas como en empresas privadas, por lo que debe ser consciente de las dificultades que encierra la utilización de los métodos numéricos contemplados en el programa de la asignatura.

Los contenidos a abordar se definen brevemente en el plan de estudios como: Errores. Algoritmos. Interpolación polinomial. Resolución numérica de ecuaciones y sistemas no lineales. Resolución numérica de sistemas de ecuaciones lineales. Derivación e integración numérica. Métodos numéricos para las ecuaciones diferenciales ordinarias y las ecuaciones en derivadas parciales, que concretaremos en los siguientes objetivos:

- Conocer las características y limitaciones de los métodos constructivos, la evolución de los errores en los cálculos numéricos y el esquema general de construcción y aplicación de los métodos iterativos.
- Conocer y saber aplicar en situaciones concretas los métodos numéricos fundamentales de resolución aproximada de ecuaciones y sistemas no lineales, estimando los errores cometidos.
- Conocer y saber aplicar los principales métodos directos e iterativos de resolución de sistemas lineales.
- Conocer y saber aplicar en situaciones concretas los problemas de interpolación y aproximación de funciones, así como los principales métodos

de interpolación por polinomios y de aproximación en espacios prehilbertianos. Obtener los polinomios de interpolación de diversas formas y aplicarlos para aproximar funciones y determinar la mejor aproximación por mínimos cuadrados continua o discreta de una función dada.

- Saber utilizar los polinomios de interpolación para obtener métodos de integración y derivación numérica de tipo interpolatorio. Y utilizar estos métodos para aproximar integrales y derivadas, estimando los errores de los mismos.

- Conocer los conceptos, resultados fundamentales y limitaciones de los métodos numéricos de un paso y de los métodos lineales multipaso, para la integración numérica de problemas de valor inicial, para ecuaciones diferenciales ordinarias. En particular, conocer y saber aplicar, en situaciones concretas, los métodos de Euler, Taylor, Runge-Kutta y Adams.

- Conocer y saber aplicar los métodos en diferencias finitas a la resolución numérica de problemas de contorno para ecuaciones diferenciales ordinarias y en derivadas parciales, en particular para la ecuación del calor, de ondas y de Laplace.

Los contenidos correspondientes a estos objetivos son desarrollados en los siete capítulos en que se ha dividido esta memoria, que van seguidos de problemas tipo resueltos (un total de 62) y de otros propuestos (68), con objeto de que el alumno pueda probar su destreza y el aprendizaje de las técnicas básicas desarrolladas en cada capítulo. La asignatura puede adaptarse fácilmente a la metodología propia del EEES, mediante la realización individual o en grupos de algunos de los trabajos propuestos y de las prácticas de la asignatura, en las que realizamos algoritmos y programas propios para los métodos fundamentales desarrollados, utilizando el software libre Maxima bajo la interface wxMaxima y el manual que hemos preparado para este fin. No obstante, hemos realizado e incluido algunos programas y funciones sencillas de Maxima para resolver algunos de los problemas tratados (utilizando las versiones 5.28.0-2 de Maxima y 12.04.0 de wxMaxima), si bien el tratamiento más completo viene en el manual de prácticas de la asignatura (referencia [14] de las citadas al final).

Aunque la bibliografía en esta materia es hoy día muy abundante, tanto en inglés como en español, me he restringido tan sólo a la realmente utilizada para preparar este manual y que se cita al final del mismo, espero que su consulta pueda servir al alumno para profundizar más en los temas tratados y para abordar con éxito algunos de los trabajos que se le puedan proponer

a lo largo del curso. Asimismo, en cuanto a la notación utilizada y resultados contemplados en los distintos capítulos, cabe destacar las referencias [7], [10], [12], [2], [5] y [9], en las que podrá ampliar conocimientos y completar algunas de las demostraciones omitidas.

Asimismo, pensamos que este manual puede ser útil también a estudiantes de otros grados de ingeniería o de ciencias cuyos planes de estudio incorporen contenidos de introducción a los métodos numéricos.

Cartagena, marzo de 2016

Capítulo 1

Cálculo Numérico: generalidades

1.1. Introducción: métodos de cálculo numérico

En este tema estudiamos el concepto de Cálculo Numérico, el esquema general de construcción y aplicación de métodos numéricos, los errores inherentes a los mismos, así como el diseño de algoritmos y diagramas de flujo.

Las leyes que rigen buena parte de los problemas estudiados en esta titulación, como transmisión de calor, dinámica de sistemas mecánicos, planificación y control de trayectorias de un robot, análisis de circuitos eléctricos, dinámica de fluidos y medios continuos, deformaciones de sólidos, propagacián de ondas, cálculo de estructuras, etc., se modelizan matemáticamente y a menudo acaban traduciéndose en sistemas de ecuaciones lineales o no, problemas de valor inicial o de contorno para ecuaciones diferenciales ordinarias o en derivadas parciales, por lo que no cabe duda de que el conocimiento de estos tópicos, al menos en sus aspectos más básicos y elementales, resulta imprescindible en la formación del ingeniero. Ahora bien, como es sabido pocos son los problemas señalados que pueden abordarse únicamente con las herramientas matemáticas vistas hasta este momento, por ello se hace necesario introducir unos conocimiento mínimos de Cálculo Numérico. En frase de **Henrici**, podríamos definir el **Cálculo Numérico** como **"la teoría de los métodos constructivos en Análisis Matemático"**. Se hace especial énfasis en la palabra **constructivo** ya que, propuesto un problema matemático, además de estudiar la existencia de solución, hay que dar un procedimiento para calcularla de modo explícito. Esta solución se construirá mediante **algoritmos**, entendiendo por tal una especificación no

ambigua de operaciones aritméticas en un orden prefijado. La aparición de los ordenadores y su creciente potencia de cálculo ha potenciado el uso y desarrollo de métodos numéricos para resolver multitud de problemas y ha hecho posible abordar problemas tan complejos como el de la predicción meteorológica. Hasta tal punto ha producido esto un cambio en la situación que el estudio de los métodos numéricos lleva camino de convertirse en una de las ramas más importantes de las mátematicas y, desde luego una de las de más actualidad. Podemos decir que un **método numérico** de resolución de un determinado problema es **"un conjunto de reglas que permite la obtención mediante un número finito de operaciones elementales, de un resultado que se aproxima de alguna manera a la solución del problema en cuestión"**. Suelen proporcionar soluciones tan próximas como se quiera (sólo en teoría puesto que al realizar los cálculos se cometen errores debido a que los ordenadores tienen una precisión limitada) a la solución exacta, pero en general no se obtiene esta. Este hecho lo pone de manifiesto **Ortega** cuando define el **Cálculo Numérico** como **"el estudio de los métodos y procedimientos usados para obtener soluciones aproximadas a problemas matemáticos"**.

Por lo dicho antes, dos serán los objetivos esenciales de esta disciplina: 1º) Dado un problema matemático, encontrar algoritmos, a veces llamados métodos numéricos, que bajo ciertas condiciones permiten obtener una solución aproximada del problema propuesto, y 2º) Analizar las condiciones bajo las cuales la solución del algoritmo es próxima a la verdadera, estimando los errores en la solución aproximada. Se trata pues de encontrar métodos aproximados para resolver todo tipo de problemas matemáticos y analizar los errores producidos en estas aproximaciones.

1.2. Esquema general de construcción y aplicación de métodos numéricos

En la construcción y aplicación de métodos numéricos para resolver un determinado problema matemático se suelen realizar los siguientes pasos:

1. Sustituir el problema inicial por un algoritmo de cálculo, que contiene un parámetro n.

2. Probar la convergencia del algoritmo, es decir asegurar que las aproximaciones, x_n, a la solución, x, son tan próximas como se desee; estimando la rapidez o velocidad de convergencia.

3. Otro requisito sobre los métodos numéricos es el de que sean estables, que significa, hablando coloquialmente, que pequeñas modificaciones en los datos no ocasionen fuertes cambios en el resultado final; o de otra forma si los cálculos no se efectúan con exactitud (debido a los redondeos) no obstante se tiene convergencia a la solución.

4. Realizar el organigrama y/o el programa del algoritmo en cuestión en un adecuado lenguaje de programación.

Criterios habitualmente utilizados para la selección de un algoritmo, entre otros posibles, para obtener la solución aproximada de un determinado problema son los siguientes:

- Elegir aquel método que minimice los errores.
- Elegir aquel método que minimice el número de operaciones efectuadas, directamente relacionado con el costo computacional.
- Elegir aquel método que minimice la memoria requerida para almacenar datos, cada vez menos importante debido a las grandes capacidades actuales de los medios de cálculo a nuestra disposición.

O simplemente, podríamos adoptar como criterio más adecuado el de **seleccionar aquel método que produciendo errores dentro de márgenes admisibles, necesite el menor costo computacional**.

1.2.1. Ejemplos que ilustran estos criterios:

- Comparar el costo computacional de diferentes formas de evaluar un polinomio: algoritmo de Horner frente a otro.

  Dado el polinomio

  $$p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$$

  con $a_i \in \mathbb{R}$ y $a_0 \neq 0$, para evaluarlo es útil escribirlo en la forma anidada

  $$p(x) = a_n + x(a_{n-1} + x(a_{n-2} + \cdots + x(a_0)\dots))$$

  que conduce al algoritmo

  $$p_0(x) = a_0$$
  $$p_k(x) = a_k + xp_{k-1}(x) \ (k = 1, 2, \dots, n)$$

resultando que $p_n(x) = p(x)$; se conoce como **algoritmo de Horner** o división sintética; es fácil comprobar que para evaluar un polinomio de grado n se requieren por este método un total de $2n$ operaciones (n sumas y n multiplicaciones), en tanto que por el método ordinario serían necesarias $3n-1$ operaciones ($2n-1$ multiplicaciones y n sumas).

- Veamos algunos ejemplos de cálculos inestables:

  a) Calcular el valor de la integral $y_{12} = \int_0^1 \frac{x^{12}dx}{10+x}$, mediante una fórmula de reducción que lleva a un algoritmo inestable por **crecimiento exponencial del error**.

  Supongamos que deseamos calcular la integral $\int_0^1 \frac{x^{12}dx}{10+x}$, para ello utilizamos la fórmula de reducción:

$$y_n = \int_0^1 \frac{x^n dx}{10+x} = \int_0^1 \frac{x^{n-1}(10+x-10)dx}{10+x} = \frac{1}{n} - 10\int_0^1 \frac{x^{n-1}dx}{10+x}$$

$$= \frac{1}{n} - 10y_{n-1}$$

  es decir la calculamos mediante el algoritmo

$$\begin{array}{c} y_0 = \log(11/10) \\ y_n = \frac{1}{n} - 10y_{n-1} \end{array}$$

  Pues bien, realizando los cálculos con wxMaxima, en doble precisión, resulta $y_0 = 0{,}095310179804325$ y obtenemos para $y_{12} = -0{,}55790049840635$ mediante este algoritmo de reducción, mientras que haciendo la integral directamente con el mismo programa se obtiene $y_{12} = 0{,}0068359375$, que muestra la inestabilidad del algoritmo utilizado para el cálculo en punto flotante. Para ilustrar la inestabilidad de este algoritmo de cálculo, supongamos que al calcular y_0 cometemos un error e_0 de modo que calcularemos $\bar{y}_0 = y_0 + e_0$ y en el paso enésimo tendremos

$$\bar{y}_n = \frac{1}{n} - 10\bar{y}_{n-1}$$

  y llamando
  $e_n = \bar{y}_n - y_n = \frac{1}{n} - 10\bar{y}_{n-1} - \frac{1}{n} + 10y_{n-1} = -10(\bar{y}_{n-1} - y_{n-1}) = -10e_{n-1} = (-10)(-10)e_{n-2} = \cdots = (-10)^n e_0$

Resulta que en cada paso el error se multiplica por -10, los algoritmos de este tipo con crecimiento exponencial del error deben ser desterrados en el cálculo numérico.

b) Calcular las raíces de la ecuación polinomial (ejemplo debido a Wilkinson, 1959) $\prod_{i=1}^{20}(x-i)=0$, tras incrementar el coeficiente de x^{19} por 10^{-7}.

La ecuación de partida tiene las raíces reales $\{1,2,\ldots,20\}$ en tanto que para la ecuación perturbada $\prod_{i=1}^{20}(x-i)+10^{-7}\cdot x^{19}=0$, obtenida sumándole a la dada el término $10^{-7}x^{19}$, con wxMaxima obtenemos sus 20 raíces, de las cuales son 10 reales y otras 10 complejas, conjugadas dos a dos, que son las siguientes:

Nº raíz	Valor obtenido
1	1,0
2	2,0
3	3,0
4	4,0
5	5,0
6	5,999992560243429
7	7,000264061262213
8	7,994096812278631
9	9,112077294685991
10	9,570864847089773
11	$10{,}92126450508518 + 1{,}102220454406112\,i$
12	$10{,}92126450508518 - 1{,}102220454406112\,i$
13	$12{,}84620554309262 + 2{,}062169061213381\,i$
14	$12{,}84620554309262 - 2{,}062169061213381\,i$
15	$15{,}31474224497542 + 2{,}69861837443531\,i$
16	$15{,}31474224497542 - 2{,}69861837443531\,i$
17	$18{,}15719154663773 + 2{,}470196797035823\,i$
18	$18{,}15719154663773 - 2{,}470196797035823\,i$
19	$20{,}421948379285 + 0{,}99919961819071\,i$
20	$20{,}421948379285 - 0{,}99919961819071\,i$

que muestra la sensibilidad de las raíces con respecto a la variación de los coeficientes de la ecuación algebraica dada.

1.3. Desarrollos de Taylor y MacLaurin

Por su utilidad en el resto de la asignatura, vamos a recordar brevemente las fórmulas de Taylor para funciones reales de una o varias variables.

Teorema 1 ***(Fórmula de Taylor con resto de Lagrange).*** *Si $f \in \mathbb{C}^{(n+1)}([a,b])$, entonces para puntos cualesquiera x y c en $[a,b]$ existe un punto ξ entre c y x tal que:*

$$f(x) = P_{n,c}(x) + R_{n,c}(x)$$

donde

$$P_{n,c}(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k$$

es el polinomio de Taylor de grado n para f en c, y

$$R_{n,c}(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}$$

es el resto enésimo de Taylor de f en c, en la forma de Lagrange, donde el punto ξ puede escribirse en la forma $\xi = c + \theta(x-c)$ con $\theta \in (0,1)$.

La demostración puede verse en cualquier libro de Cálculo Infinitesimal. En particular, si $c = 0$ el desarrollo anterior se denomina de MacLaurin y queda como sigue

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k + \frac{f^{(n+1)}(\theta x)}{(n+1)!}x^{n+1}$$

con $\theta \in (0,1)$.

Algunos desarrollos notables son los siguientes:

- $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!}e^{\theta x}$.
- $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{(-1)^{k-1}x^{2k-1}}{(2k-1)!} + \frac{\sin(\theta x + (2k+1)\frac{\pi}{2})}{(2k+1)!}x^{2k+1}$.
- $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots + \frac{(-1)^{n-1}x^n}{n} + \frac{(-1)^n x^{n+1}}{(n+1)(\theta x+1)^{n+1}}$.

Para funciones reales de varias variables la fórmula de Taylor viene dada en el siguiente.

Teorema 2 *Sea $f \in \mathbb{C}^{(n+1)}(\Omega_n)$, donde Ω_n es una abierto de $\mathbb{R}^n$, entonces para puntos cualesquiera x y c en Ω_n tal que el segmento $[c,x] \subset \Omega_n$, existe un punto ξ intermedio en el segmento $[c,x]$ tal que:*

$$f(x_1, x_2, \ldots, x_n) = P_{n,c}(x_1, x_2, \ldots, x_n) + R_{n,c}(x_1, x_2, \ldots, x_n)$$

donde

$$P_{n,c}(x) = \sum_{k=0}^{n} \frac{1}{k!}\{(x_1 - c_1)\frac{\partial}{\partial x_1} + \cdots + (x_n - c_n)\frac{\partial}{\partial x_n}\}^{(k)} f(c_1, \ldots, c_n)$$

es el polinomio de Taylor de grado n para f en c, y

$$R_{n,c}(x) = \frac{1}{(n+1)!}\{(x_1 - c_1)\frac{\partial}{\partial x_1} + \cdots + (x_n - c_n)\frac{\partial}{\partial x_n}\}^{(n+1)} f(c + \theta(x - c))$$

es el resto enésimo de Taylor de f en c, con $c + \theta(x - c) = (c_1 + \theta(x_1 - c_1), \ldots, c_n + \theta(x_n - c_n))$ *y* $0 < \theta < 1$.

Observación. *El primer sumatorio del polinomio de Taylor cuando* $k = 0$ *nos da la propia* $f(c_1, \ldots, c_n)$*, para otros valores de k, los términos de la potencia* $\{\ldots\}^{(k)}$ *nos dan productos de monomios de orden k por derivadas del mismo orden de f en c.*

En particular si $c = (0, \ldots, 0)$ el desarrollo anterior se denomina de MacLaurin.

1.4. Orden y rapidez de convergencia: notaciones o y O

1.4.1. Las notaciones o y O de Landau para funciones

Sean f y g aplicaciones reales definidas en un entorno del punto x_0, decimos que f **es una o minúscula de** g **cuando** x **tiende a** x_0 (y escribimos $f(x) = o(g(x))$ cuando $x \to x_0$) si para x próximo a x_0 con $x \neq x_0$, g no se anula y $\lim_{x\to x_0} f(x)/g(x) = 0$, es decir si para todo $\varepsilon > 0$ existe un $r > 0$, tal que para todo x verificando $0 <\mid x - x_0 \mid< r$ es $\mid f(x) \mid< \varepsilon \mid g(x) \mid$; la definición viene a expresar la idea de que f(x) se vuelve despreciable frente a g(x) cuando x tiende a x_0.

A veces interesa lo que ocurre solamente cuando nos acercamos a x_0 por la derecha, por ello también decimos que f **es una o minúscula de** g **cuando** x **tiende a** x_0 **por la derecha** (y escribimos $f(x) = o(g(x))$ cuando $x \to x_0^+$) si para x próximo a x_0 por la derecha con $x \neq x_0$, g no se anula y $\lim_{x\to x_0^+} f(x)/g(x) = 0$, es decir si para todo $\varepsilon > 0$ existe un $r > 0$, tal que para todo x verificando $x_0 < x < x_0 + r$ es $\mid f(x) \mid< \varepsilon \mid g(x) \mid$.

Otra notación habitual para hablar del tamaño relativo de las funciones en las proximidades de un punto es la siguiente: decimos que f **es una O mayúscula de** g **cuando** x **tiende a** x_0 (y escribimos $f(x) = O(g(x))$ cuando $x \to x_0$) si existen constantes K y r positivas tales que para todo x verificando $\mid x - x_0 \mid< r$ es $\mid f(x) \mid\leq K \mid g(x) \mid$. De modo similar al caso anterior puede definirse la **O mayúscula cuando** x **tiende a** x_0 **por la derecha**.

Las definiciones anteriores pueden extenderse sin dificultad al caso en que $x_0 = \pm\infty$.

En ocasiones lo que interesa es comparar una función $f(x)$ con monomios $g(x) = (x - x_0)^m$ cuando $x \to x_0$, así hablamos de que $f(x) = o((x - x_0)^m)$ cuando $x \to x_0$ para expresar que $f(x)$ es un infinitésimo de orden superior a $(x-x_0)^m$ cuando $x \to x_0$, es decir $f(x)$ **converge a cero más rápidamente que lo hace** $(x - x_0)^m$ **cuando** $x \to x_0$. En algunas ocasiones $f(h)$ expresa el error que se comete en un método numérico que depende de un parámetro de discretización positivo h, en este contexto decimos que $f(h) = O(h^m)$ cuando $h \to 0^+$ si existen constantes K y r positivas tales que para todo h verificando $0 < h < r$ es $| f(h) | \leq Kh^m$, lo que viene a expresar que $f(h)$ **converge a** 0^+ **al menos tan rápido como** h^m.

1.4.2. Las notaciones o y O de Landau para sucesiones: órdenes de convergencia

Dadas dos sucesiones $\{x_n\}$ e $\{y_n\}$, tal que $y_n \neq 0$ para toda n, se dice que $x_n = o(y_n)$ si $\lim_{n\to\infty} x_n/y_n = 0$; asimismo, se dice que $x_n = O(y_n)$ si existen constantes positivas K y r tales que $| x_n | \leq K | y_n |$ para toda $n \geq r$, en caso de ser $y_n \neq 0$ para toda n, esto significa que $| x_n/y_n |$ permanece acotada por K cuando $n \to \infty$. Si ambas sucesiones tienden a cero, en el primer caso se dice que la sucesión x_n **converge a cero más rápidamente** que lo hace la y_n, y en el segundo que **converge al menos tan rápidamente** como esta.

Dada una sucesión de números reales $\{x_n\}$ convergente al número real x, se dice que la convergencia es:

- **lineal** si existe una constante positiva $c < 1$ y un entero n_0 tal que $| x_{n+1} - x | \leq c | x_n - x |$ para todo $n \geq n_0$.
- **superlineal** si existe una sucesión convergente a cero ξ_n y un entero n_0 tal que $| x_{n+1} - x | \leq \xi_n | x_n - x |$ para todo $n \geq n_0$.
- **de orden, al menos,** $p > 1$ si existen dos constantes K (no necesariamente menor que 1) y n_0 tal que $| x_{n+1} - x | \leq K | x_n - x |^p$ para todo $n \geq n_0$. Si $p = 2$, la convergencia se dice, al menos, cuadrática; si $p = 3$ cúbica, etc.

1.5. Errores en los métodos numéricos

La aplicación de métodos numéricos nos lleva a la consideración de los errores inherentes a los mismos, que son básicamente los que siguen:

- **Errores en los datos iniciales**, por ejemplo si son resultado de una medida con algún instrumento.
- **Errores de redondeo**, debidos al hecho de que el ordenador maneja sólo un número finito de cifras significativas o dígitos.
- **Errores de truncatura o discretización**, que provienen de sustituir un problema contínuo por otro discreto, por ejemplo una serie por una suma finita, una derivada por un cociente incremental, o una integral definida por una suma de un número finito de términos, etc.

En Cálculo Numérico se estudian básicamente los errores de los dos últimos tipos, sobre todo los de truncatura o discretización como veremos en las lecciones siguientes.

Obtenida una aproximación $\bar{x}$ del verdadero valor x de una determinada magnitud, se llama **error absoluto** $x-\bar{x}$, si $\bar{x} > x$ se dice que la aproximación es por **exceso** y si $\bar{x} < x$ se dice **por defecto**, generalmente interesa el error sin signo o sea el $\mid e \mid$, además si $x \neq 0$ se define el **error relativo** por $\mid e \mid / \mid x \mid$, si este último se multiplica por 100 se tiene el **error porcentual**. Normalmente, como no es usual conocer los errores con exactitud, trataremos de obtener cotas de estos errores.

Aunque en los siguientes capítulos no analizaremos con detalle los errores de redondeo y su propagación, conviene que reparemos en algunas observaciones al respecto. En primer lugar, señalemos que no todo número real puede representarse de modo exacto en un ordenador digital, de hecho ni tan siquiera todos los números decimales con un número finito de cifras, como por ejemplo 0,1, pues en base 2 posee infinitos dígitos, de hecho es $0{,}1 = 0{,}00011001100110011..._{(2}$. Por ello, al ser almacenado el número 0,1 en un ordenador digital sufrirá un redondeo, es decir que no sólo lo sufren los irracionales o los racionales con infinitas cifras decimales.

Para hacernos una idea es suficiente con trabajar en base 10. Como es sabido todo número real no nulo x puede expresarse de modo único en la forma $x = \pm m \cdot 10^q$ con $0{,}1 \leq m < 1$ y q entero, a m se le denomina mantisa y a q exponente, para almacenar dicho número en el ordenador, debemos guardar su signo, la mantisa y el exponente con su signo, pero para guardar m y q sólo podemos utilizar un número finito de dígitos, si por ejemplo q sólo pudiera tener dos dígitos en base 10 el intervalo de los números representables estaría contenido en el dado por

$$10^{-100} \leq \mid x \mid < 10^{99}$$

pero dentro de este intervalo, por las razones antes expuestas, no todos los números pueden ser representados con exactitud, pues si por ejemplo sólo

pueden ser representados k dígitos para almacenar m, dado x cualquiera del intervalo anterior, quedará representado como

$$\bar{x} = \pm \bar{m} \cdot 10^q$$

con $\bar{m}$ obtenida por redondeo (lo más usual) o corte tras el k-ésimo dígito.

El **error absoluto** $| e | = | x - \bar{x} |$ está acotado por $\frac{1}{2} 10^{-k} \cdot 10^q$ (en redondeo) y por $10^{-k} \cdot 10^q$ (en corte). Y el **error relativo** en x, supuesto este no nulo, está acotado por $\frac{1}{2} \cdot 10^{-k+1}$ (en redondeo) y por 10^{-k+1} en corte.

En general, los ordenadores trabajan en aritmética de punto **flotante**, con un número fijo k de dígitos; esta aritmética tiene propiedades algo distintas de las habituales, por ejemplo la suma deja de ser asociativa, hay números no nulos que al ser sumados a otro no alteran la suma, se produce la pérdida de cifras significativas al hacer la diferencia de números próximos, etc. Veremos diversos ejemplos en las prácticas de la asignatura al trabajar en aritmética de redondeo y calcularemos el denominado **épsilon de máquina** ϵ, que definiremos como el número positivo más pequeño que al sumar a otro número a verifica que $\epsilon + a > a$, de manera que todo número positivo ϵ' menor que ϵ verifica que $\epsilon' + a = a$. Una cifra decimal, la a_r correspondiente a la potencia de 10^{-r}, diremos que es **correcta** si el módulo del error absoluto es menor que $5 \cdot 10^{-(r+1)}$.

1.6. Algoritmos y diagramas de flujo

- Podemos decir que un **algoritmo** es un procedimiento que describe, sin ambigüedad alguna, una sucesión finita de pasos a realizar en un orden específico para resolver un determinado problema.

- Un **organigrama o diagrama de flujo** es una representación gráfica del algoritmo, que nos da los pasos del algoritmo y el flujo de control entre los diferentes pasos. Los distintos tipos de operaciones se indican por distintos tipos de cajas, y los flujos de control se indican por líneas dirigidas entre las cajas.

El esquema general de un organigrama está constituido por tres bloques principales, correspondientes a los tres tipos de operaciones siguientes:

- **Operaciones de entrada**. Son utilizadas para la entrada de datos desde el exterior a las unidades de almacenamiento.

- **Operaciones de proceso**. Se emplean para la manipulación de los datos anteriormente almacenados, de donde se obtendrán unos resultados que serán asimismo almacenados.
- **Operaciones de salida**. Son aquellas que comunican al usuario unos resultados, extraídos de las unidades de almacenamiento.

Un subconjunto de un programa tiene una estructura:

- **Repetitiva**, si consta de una secuencia lógica principio, una secuencia lógica fin, ejecutadas una vez en el conjunto y un subconjunto repetitivo ejecutado n veces.
- **Alternativa**, si consta de una secuencia lógica principio, una secuencia lógica fin, ejecutadas una vez en el conjunto y dos o más subconjuntos mutuamente excluyentes, de manera que si se realiza uno de ellos no se realizan los restantes.
- **Compleja**, si consta de varias estructuras elementales, alternativas o repetitivas. Si sólo consta de estructuras repetitivas, se dice repetitivo complejo; si sólo consta de estructuras alternativas se dice alternativo complejo y si comprende de ambas se dice complejo mixto.

Cada una de estas operaciones se representan gráficamente por símbolos, por ejemplo por círculos para iniciar o parar, por rectángulos para cálculo o proceso diferente de una decisión, por rombos para las decisiones, por trapecios para la entrada de datos, se conectan las diversas partes del programa mediante una línea de flujo, etc. A veces representamos los algoritmos mediante un seudocódigo, a partir del cual se puede realizar el programa correspondiente en algún lenguaje de programación, como señalamos en el prólogo nosotros utilizamos Maxima bajo la interface wxMaxima.

Realizar los organigramas de los programas siguientes:

- Del que calcula las normas $\| \cdot \|_1$ y $\| \cdot \|_2$ de un vector de $\mathbb{R}^n$.

  Recordemos que dado $x \in \mathbb{R}^n$, se definen las normas $\| x \|_1 = \sum_{i=1}^{n} | x_i |$ y la $\| x \|_2 = (\sum_{i=1}^{n} | x_i |^2)^{\frac{1}{2}}$.

  El organigrama para el primero, en seudocódigo, sería el siguiente:

  Principio: Leer $n, x_i (i = 1, 2, \dots, n)$
  Proceso:(realizar el sumatorio)
  $Norma1 = 0$

desde $i = 1$ hasta n, hacer
$Norma1 = Norma1 + \mid x_i \mid$
Fin: Escribir $Norma1$.

El otro es similar pero haciendo la raíz cuadrada antes de salir.

- Del que ordena de mayor a menor dos números reales dados.

Principio: Leer x_1, x_2
Proceso:
Si $x_1 - x_2 \geq 0$ ir al fin
si no hacer $var = x_2, x_2 = x_1, x_1 = var$
Fin: Escribir x_1, x_2.

1.7. Problemas resueltos

1. ¿Es $\cot x = o(x^{-1})$ cuando $x \to 0$?.

   Solución. Para responder a esta cuestión es suficiente con realizar $\lim_{x\to 0} \frac{\cot x}{1/x} = \lim_{x\to 0} x \cot x = \lim_{x\to 0} x \frac{\cos x}{\sin x} = 1$, por tanto no se verifica que $\cot x$ sea una $o(x^{-1})$.

2. Probar que dada la sucesión $\{x_n\}$, entonces $x_n = x + o(1)$ si y sólo si $lim_{n\to\infty} x_n = x$.

   Solución. En efecto, $x_n = x + o(1)$ si y sólo si $x_n - x = o(1)$ si y sólo si $\lim_{n\to\infty}(x_n - x) = 0$ o sea si y sólo si $\lim_{n\to\infty} x_n = x$.

3. ¿Averiguar, utilizando el desarrollo de Maclaurin, si $e^x - 1 = O(x^2)$ cuando $x \to 0$?.

   Solución. El desarrollo en serie de MacLaurin, con resto de Lagrange, puede escribirse como $e^x - 1 = x + \frac{x^2}{2!} e^{\theta x}$, luego $e^x - 1$ no es una $O(x^2)$ sino una $O(x)$ cuando $x \to 0$?.

4. Utilizando el desarrollo de Maclaurin de la función $\sin x$, probar que $\sin x = x - \frac{x^3}{6} + O(x^5)$ cuando $x \to 0$.

   Solución. Es evidente a partir del desarrollo de MacLaurin de la función $\sin x$.

5. Probar que $\frac{n+1}{n^3} = o(\frac{1}{n})$ cuando $n \to \infty$.

   Solución. Basta ver que $\lim_{n\to\infty} \frac{\frac{n+1}{n^3}}{\frac{1}{n}} = \lim_{n\to\infty} \frac{n^2+n}{n^3} = 0$.

6. Probar que si $f \in \mathbb{C}^{(n)}([a,b])$ se anula en $n+1$ puntos distintos $x_0 < x_1 < \ldots < x_n$ de dicho intervalo, entonces existe un $\xi \in (x_0, x_n)$ tal que $f^{(n)}(\xi) = 0$.

   Solución. Debido a la hipótesis de ser f n veces derivable con derivada continua y anularse en cada x_i, por el teorema de Rolle, la derivada f' se anula en, al menos, n puntos ξ_i, cada uno de ellos intermedio entre x_{i-1} y x_i, aplicándole a f' el mismo teorema la f'' se anula en, al menos, $n-1$ puntos intermedios entre los ξ_i, y finalmente la $f^{(n)}$ se anula, al menos, en un punto ξ intermedio entre x_0 y x_n.

1.8. Aritmética exacta versus aritmética de redondeo con Maxima

Maxima es un sistema para la manipulación de expresiones simbólicas y numéricas, incluyendo diferenciación, integración, desarrollos en series de Taylor, transformadas de Laplace, ecuaciones diferenciales ordinarias, sistemas de ecuaciones lineales, vectores, matrices y tensores. También realiza representaciones gráficas de funciones y datos en dos y tres dimensiones. Se trata de un software libre disponible en http://maxima.sourceforge.net/es/.

Puede trabajarse en línea de comandos, pero hay dos opciones gráficas, nosotros trabajaremos con la opción wxMaxima, que es bastante más agradable, y suele venir con la distribución de Maxima.

Maxima produce resultados con alta precisión usando fracciones exactas y representaciones con aritmética de coma flotante arbitraria. Por defecto, hace las operaciones encomendadas de forma exacta, por ejemplo la suma de fracciones devuelve otra fracción y lo mismo la raíz cuadrada u otras funciones cuyo resultado no sea un entero, a no ser que se le pida mediante **float(número)** o **número, numer**, que dan la expresión decimal de número con 16 dígitos o también con **bfloat(numero)** que da la expresión decimal larga de número acabada con b seguido de un número n, que significa multiplicar por 10^n, en este caso el número de cifras se precisa previamente con el comando **fpprec:m**, por defecto es **fpprec:16**. Pero trabajar en aritmética exacta no es conveniente en Cálculo Numérico, pues el tiempo empleado (que se le pide mediante la instrucción **showtime:true**) en los cálculos aumenta considerablemente, no siendo proporcional al número de operaciones efectuadas y en muchas ocasiones, aunque Maxima calcule las operaciones encomendadas, no llega a mostrar los resultados. Veamos lo que acabamos de afirmar en los siguientes ejemplos, en los que se calcula la suma de los inversos de los cuadrados de los $n = 100, 1000, 10000$ primeros números na-

turales en aritmética exacta y se muestra el tiempo de cálculo empleado en realizar dicha suma:

```
(%i1) numer:false;
```

(%o1) *false*

```
(%i2) showtime:true$ sum (1/i^2,i,1,100);
```

Evaluation took 0,0000 *seconds* (0,0000 *elapsed*)
Evaluation took 0,0000 *seconds*(0,0000 *elapsed*)

(%o3)
$$\frac{1589508694133037873112297928 5\cdots 37098598894328348038181310903699 01}{97218614443438103058965797 66\cdots 57462417827203547055179861652480 00}$$

```
(%i4) sum (1/i^2,i,1,1000);
```

Evaluation took 0,0200 *seconds* (0,0200 *elapsed*)

(%o4) $\frac{83545938483149689478187[820 digits]58699094240207812766449}{50820720104325812617835[820 digits]01453118476390400000000}$

```
(%i5) sum (1/i^2,i,1,10000);
```

Evaluation took 0,3500 *seconds*(0,3500 *elapsed*)

(%o5) $\frac{54714423173933343999582[8648 digits]71491756496677000005717}{33264402841837255560349[8648 digits]95863724851200000000000}$

```
(%i6) sum (1/i^2,i,1,100000);
```

Evaluation took 20.5300 seconds (20.5300 elapsed) « ¡Expresión excesivamente larga para ser mostrada! »

En general, los ordenadores trabajan en aritmética de redondeo, también llamada de punto **flotante**, con un número fijo k de dígitos; esta aritmética tiene propiedades algo distintas de las habituales, por ejemplo la suma deja de ser asociativa, hay números no nulos que al ser sumados a otro no alteran la suma, se produce la pérdida de cifras significativas al hacer la diferencia de números próximos, etc.; pero se suelen obtener resultados satisfactorios en un tiempo razonable, en general muchísimo menor que en la aritmética exacta y los tiempos empleados tienden a ser proporcionales al número de operaciones realizadas. Recordemos también que por defecto Maxima trabaja en aritmética exacta, de manera que si queremos que lo haga en aritmética de redondeo podemos declararlo mediante **numer:true**. Veamos diversos ejemplos con el anterior sumatorio:

```
(%i7) numer:true$ sum (1/i^2,i,1,100);
```

Evaluation took 0,0000 *seconds*(0,0000 *elapsed*)
Evaluation took 0,0000 *seconds*(0,0000 *elapsed*)
(%o8) 1,634983900184892

```
(%i9)  sum (1/i^2,i,1,1000);
```

Evaluation took 0,0100 *seconds*(0,0100 *elapsed*)
(%o9) 1,643934566681561

```
(%i10) sum (1/i^2,i,1,10000);
```

Evaluation took 0,0500 *seconds*(0,0500 *elapsed*)
(%o10) 1,644834071848065

```
(%i11) sum (1/i^2,i,1,100000);
```

Evaluation took 0,3700 *seconds* (0,3700 *elapsed*)
(%o11) 1,644924066898243

Se observa que la aritmética de redondeo es mucho más rápida que la exacta. Seguidamente, vamos a aumentar el número de términos en el sumatorio anterior a $1, 2, 4$ u 8 millones para ver como el tiempo de cálculo empleado tiende a ser proporcional al número de operaciones realizadas, primero la ponemos en punto flotante de 16 cifras, luego mostramos el tiempo de cálculo y realizamos los sumatorios pedidos (se trata del ejercicio nº 14 de los propuestos al final de este tema):

```
(%i12) numer:true;
```

Evaluation took 0,0000 *seconds*(0,0000 *elapsed*)
(%o12) *true*

```
(%i14) showtime:true;
```

Evaluation took 0,0000 *seconds*(0,0000 *elapsed*)
(%o14) *true*

```
(%i15) sum (1/i^2,i,1,1000000);
```

Evaluation took 3,3980 *seconds*(3,4000 *elapsed*)
(%o15) 1,644933066848771

```
(%i16) sum (1/i^2,i,1,2000000);
```

Evaluation took 6,5300 *seconds*(6,5400 *elapsed*)
(%o16) 1,644933566848396

```
(%i17) sum (1/i^2,i,1,4000000);
```

Evaluation took 12,5300 *seconds*(12,5400 *elapsed*)
(%o17) 1,644933816848455

```
(%i18) sum (1/i^2,i,1,8000000);
```

Evaluation took 24,6100 *seconds*(24,6200 *elapsed*)
(%o18) 1,644933941848658

Se observa el crecimiento lineal del tiempo de CPU en la aritmética de redondeo (lo que se pone de manifiesto con la gráfica discreta de tiempos empleados frente al número de operaciones), no ocurre lo mismo en la aritmética exacta.

```
(%i19) wxplot2d([discrete, [[1000000, 3.398],[2000000, 6.53],
       [4000000,12.53], [8000000,24.61]]],[style,points]);
```

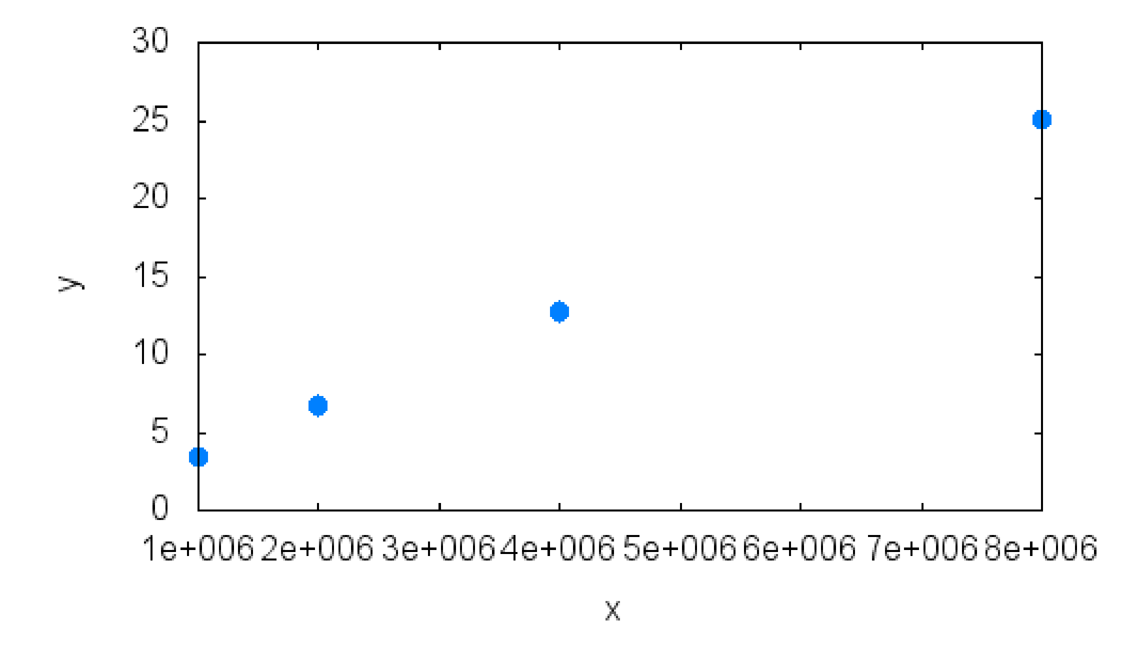


Figura 1.1: Crecimiento lineal del tiempo de CPU en aritmética de redondeo

En el comando **wxplot2d([discrete, [[1000000, 3.398],[2000000, 6.53], [4000000,12.53], [8000000,24.61]]],[style,points])**, el prefijo wx hace que lo dibuje en la pantalla, el término discrete le informa del tipo de dibujo, los puntos se introducen por pares y style le informa que es un gráfico de puntos.

1.8.1. Épsilon de máquina

Cuando trabajamos en aritmética de redondeo, conviene recordar que denominamos **épsilon de máquina** ϵ, al número positivo más pequeño que al ser sumado a otro número cualquiera a verifica que $\epsilon + a > a$, de manera que todo número positivo ϵ' menor que ϵ verifica que $\epsilon' + a = a$. Se trata pues de hallar el primer número positivo ϵ tal $1 + \epsilon > 1$, en base 2 será de la forma $0{,}0000000 \cdots 01_{(2}$, luego una potencia de exponente negativo de 2, hemos de hallar pues el primer número n tal que $1 + 2^{-n} = 1$, en aritmética de punto flotante, en cuyo caso el cero de máquina será $2^{-(n-1)} = 2^{(-n+1)}$ y puede obtenerse, por ejemplo, con el programa:

```
(%i21) kill(all)$ fpprec:20$ n:0$ while (1.0+2^(-n)>1.0)
       do (print(" Para n = ",n, "es 1.0+2^(-",n,") = ",
       bfloat(1+2^(-n))," > 1"),n:n+1)$ print (" El épsilon de
       máquina es 2^(",-n+1,") = ",float(2^(-n+1)))$
```

Salida:
Para n = 0 es $1{,}0 + 2^{-0} = 2{,}0b0 > 1$
Para n = 1 es $1{,}0 + 2^{-1} = 1{,}5b0 > 1$
Para n = 2 es $1{,}0 + 2^{-2} = 1{,}25b0 > 1$
Para n = 3 es $1{,}0 + 2^{-3} = 1{,}125b0 > 1$
Para n = 4 es $1{,}0 + 2^{-4} = 1{,}0625b0 > 1$
Para n = 5 es $1{,}0 + 2^{-5} = 1{,}03125b0 > 1$
Para n = 6 es $1{,}0 + 2^{-6} = 1{,}015625b0 > 1$
Para n = 7 es $1{,}0 + 2^{-7} = 1{,}0078125b0 > 1$
Para n = 8 es $1{,}0 + 2^{-8} = 1{,}00390625b0 > 1$
Para n = 9 es $1{,}0 + 2^{-9} = 1{,}001953125b0 > 1$
Para n = 10 es $1{,}0 + 2^{-10} = 1{,}0009765625b0 > 1$
Para n = 11 es $1{,}0 + 2^{-11} = 1{,}00048828125b0 > 1$
Para n = 12 es $1{,}0 + 2^{-12} = 1{,}000244140625b0 > 1$
Para n = 13 es $1{,}0 + 2^{-13} = 1{,}0001220703125b0 > 1$
Para n = 14 es $1{,}0 + 2^{-14} = 1{,}00006103515625b0 > 1$
Para n = 15 es $1{,}0 + 2^{-15} = 1{,}000030517578125b0 > 1$
Para n = 16 es $1{,}0 + 2^{-16} = 1{,}0000152587890625b0 > 1$
Para n = 17 es $1{,}0 + 2^{-17} = 1{,}00000762939453125b0 > 1$
Para n = 18 es $1{,}0 + 2^{-18} = 1{,}000003814697265625b0 > 1$
Para n = 19 es $1{,}0 + 2^{-19} = 1{,}0000019073486328125b0 > 1$
Para n = 20 es $1{,}0 + 2^{-20} = 1{,}0000009536743164063b0 > 1$
Para n = 21 es $1{,}0 + 2^{-21} = 1{,}0000004768371582031b0 > 1$
Para n = 22 es $1{,}0 + 2^{-22} = 1{,}0000002384185791016b0 > 1$
Para n = 23 es $1{,}0 + 2^{-23} = 1{,}0000001192092895508b0 > 1$

Para n $= 24$ es $1{,}0 + 2^{-24} = 1{,}0000000596046447754b0 > 1$
Para n $= 25$ es $1{,}0 + 2^{-25} = 1{,}0000000298023223877b0 > 1$
Para n $= 26$ es $1{,}0 + 2^{-26} = 1{,}0000000149011611939b0 > 1$
Para n $= 27$ es $1{,}0 + 2^{-27} = 1{,}0000000074505805969b0 > 1$
Para n $= 28$ es $1{,}0 + 2^{-28} = 1{,}0000000037252902985b0 > 1$
Para n $= 29$ es $1{,}0 + 2^{-29} = 1{,}0000000018626451492b0 > 1$
Para n $= 30$ es $1{,}0 + 2^{-30} = 1{,}0000000009313225746b0 > 1$
Para n $= 31$ es $1{,}0 + 2^{-31} = 1{,}0000000004656612873b0 > 1$
Para n $= 32$ es $1{,}0 + 2^{-32} = 1{,}0000000002328306437b0 > 1$
Para n $= 33$ es $1{,}0 + 2^{-33} = 1{,}0000000001164153218b0 > 1$
Para n $= 34$ es $1{,}0 + 2^{-34} = 1{,}0000000000582076609b0 > 1$
Para n $= 35$ es $1{,}0 + 2^{-35} = 1{,}0000000000291038305b0 > 1$
Para n $= 36$ es $1{,}0 + 2^{-36} = 1{,}0000000000145519152b0 > 1$
Para n $= 37$ es $1{,}0 + 2^{-37} = 1{,}0000000000072759576b0 > 1$
Para n $= 38$ es $1{,}0 + 2^{-38} = 1{,}0000000000036379788b0 > 1$
Para n $= 39$ es $1{,}0 + 2^{-39} = 1{,}0000000000018189894b0 > 1$
Para n $= 40$ es $1{,}0 + 2^{-40} = 1{,}0000000000009094947b0 > 1$
Para n $= 41$ es $1{,}0 + 2^{-41} = 1{,}0000000000004547474b0 > 1$
Para n $= 42$ es $1{,}0 + 2^{-42} = 1{,}0000000000002273737b0 > 1$
Para n $= 43$ es $1{,}0 + 2^{-43} = 1{,}0000000000001136868b0 > 1$
Para n $= 44$ es $1{,}0 + 2^{-44} = 1{,}0000000000000568434b0 > 1$
Para n $= 45$ es $1{,}0 + 2^{-45} = 1{,}0000000000000284217b0 > 1$
Para n $= 46$ es $1{,}0 + 2^{-46} = 1{,}0000000000000142109b0 > 1$
Para n $= 47$ es $1{,}0 + 2^{-47} = 1{,}0000000000000071054b0 > 1$
Para n $= 48$ es $1{,}0 + 2^{-48} = 1{,}0000000000000035527b0 > 1$
Para n $= 49$ es $1{,}0 + 2^{-49} = 1{,}0000000000000017764b0 > 1$
Para n $= 50$ es $1{,}0 + 2^{-50} = 1{,}0000000000000008882b0 > 1$
Para n $= 51$ es $1{,}0 + 2^{-51} = 1{,}0000000000000004441b0 > 1$
Para n $= 52$ es $1{,}0 + 2^{-52} = 1{,}000000000000000222b0 > 1$
El épsilon de máquina es $2^{-52} = 2{,}22044604925031 31b - 16$.

Notemos que al poner 1.0 en el programa anterior, este fuerza a Maxima a trabajar en aritmética de redondeo. Otra forma de obtener el épsilon de máquina la da el programa:

```
(%i5)  kill(all)$ epsilon:1.0$
       while ((1+epsilon/2)>1) do(
           epsilon:epsilon/2)$
       print("El épsilon de máquina de Maxima: ",float(epsilon))$
```

Salida: El épsilon de máquina de Maxima: $2{,}2204460492503131 \cdot 10^{-16}$.

Podemos preguntar si el número hallado cumple la definición en la forma

```
(%i6) is (1+2.220446049250313l*10^-16>1);
```

(%o6) *true*

```
(%i7) is (1+(2.220446049250313l*10^-16)/2>1);
```

(%o7) *false*

1.9. Problemas y trabajos propuestos

Problemas propuestos:

El alumno debe preparar el organigrama o programa para poder resolver los siguientes problemas en las clases de prácticas con el software Maxima.

1. Obtener el desarrollo de Taylor para $f(x) = e^x$ entorno al punto $x = 2$; asimismo, obtenerlo a partir de la serie para f en torno a $x = 0$.

2. El menor valor positivo ϵ, tal que $1 + \epsilon > 1$ se denomina épsilon de máquina, este depende del hardware y de como el compilador del lenguaje de programación utilizado almacena la mantisa de un número en punto flotante. Se pide que determinéis el épsilon de máquina, en base dos, para el ordenador y el lenguaje que utilicéis.

3. Para ilustrar la pérdida de dígitos significativos por sustracción de cantidades casi iguales, calcular $\sqrt{x^2+1} - 1$ y $x - senx$ para valores de x próximos a cero. Disponer los cálculos de manera que, en ambos casos, se evite dicha pérdida.

4. Para calcular la integral $I_n = \int_0^1 x^n e^x dx (n \geq 0)$ podemos utilizar (integrando por partes) el algoritmo $I_{n+1} = e - (n+1)I_n$ con $I_0 = e - 1$. Obtener mediante este algoritmo I_{18}, I_{20}, I_{21}, I_{30}, I_{31}, etc., contradice esto el hecho de que $lim_{n\to\infty} I_n = 0$, que se puede decir de la estabilidad de este algoritmo.

Trabajos propuestos:

Para los trabajos propuestos en esta y en sucesivas lecciones, se pide el análisis teórico de los mismos, diversas aplicaciones y el algoritmo o programa computacional necesario para su implementación práctica. También deben estar bien escritos y ser presentados en clase por sus autores.

- Análisis del error en la Aritmética de punto flotante.

Capítulo 2

Resolución numérica de ecuaciones y sistemas no lineales

2.1. Introducción

Ruffini y Abel demostraron que, en general, no se puede resolver la ecuación algebraica de orden n, con $n > 4$, por radicales. Aunque hay ciertos tipos de ecuaciones de grado superior al cuarto, como por ejemplo las binomias $x^n + a = 0$ que se resuelven de esta forma. Pero esta imposibilidad no ocasiona mayores inconvenientes, pues se dispone de diversos métodos que permiten calcular (según veremos a lo largo de este tema) los valores numéricos de las raíces con tanta aproximación como se desee.

El proceso a seguir para determinar las raíces reales de una ecuación dada $f(x) = 0$ (algebraica o no) suele ser el siguiente, se comienza determinando un intervalo donde estén todas, es lo que se llama **acotación de las raíces**. Seguidamente, se subdivide este intervalo en otros tales que cada uno contenga solamente una raíz, es lo que se denomina **separación de raíces**. Finalmente, se reduce la amplitud de dichos intervalos hasta que sus extremos constituyan valores aproximados aceptables, por exceso o defecto, de la raíz contenida en su interior, en esto consiste la **aproximación de raíces**.

Algunos resultados prácticos sobre acotación de raíces de ecuaciones polinómicas se verán más adelante; ahora recordamos ciertos resultados de cálculo infinitesimal por su interés en la separación de raíces, comenzando por el teorema de Bolzano.

Teorema 3 *Sea $f : [a, b] \to \mathbb{R}$ una función real, continua y tal que toma valores opuestos en los extremos del intervalo (es decir es $f(a)f(b) < 0$);*

entonces, existe $r \in (a, b)$ *tal que* $f(r) = 0$.

Observación.- El teorema asegura que, en las condiciones anteriores, existe al menos un punto interior r en el que $f(x)$ se anula, pero no que este r sea único, puede o no serlo, aunque se pueden dar condiciones adicionales para precisar la unicidad, como en la siguiente proposición.

Proposición 1 *Si además de las condiciones del teorema 1 anterior, se supone que* f *es derivable en el intervalo abierto* (a, b) *con* $f'(x) > 0\ \forall x \in (a, b)$ *(o* $f'(x) < 0\ \forall x \in (a, b)$*), entonces la raíz* $r \in (a, b)$ *de la ecuación* $f(x) = 0$ *cuya existencia se garantiza en el teorema 1* ***es única***.

Demostración. En efecto, si es $f'(x) > 0\ \forall x \in (a, b)$ entonces f es estrictamente creciente en (a, b) por tanto no puede tener raíces distintas, para el otro caso f es estrictamente decreciente y la demostración es análoga.

Un resultado sencillo que nos permite en muchos casos acotar el error absoluto que se comete al tomar el valor aproximado α de una raíz r viene dado por la proposición siguiente.

Proposición 2 *Sea* r *una raíz de la ecuación* $f(x) = 0$*,* α *un valor aproximado de la misma, tal que* $r, \alpha \in [a, b]$*. Si* f *es derivable en* $[a, b]$ *y* $\forall x \in [a, b]$ *es* $\mid f'(x) \mid \geq m > 0$*, entonces*

$$\boxed{\mid \alpha - r \mid \leq \frac{\mid f(\alpha) \mid}{m}}$$

.

Demostración. Por ser r una raíz de la ecuación $f(x) = 0$, el teorema del valor medio permite escribir $f(\alpha) = f(\alpha) - f(r) = (\alpha - r)f'(c)$ con c intermedio entre α y r, luego $\mid f(\alpha) \mid = \mid \alpha - r \mid\mid f'(c) \mid \geq \mid \alpha - r \mid m$, de donde se sigue que $\mid \alpha - r \mid \leq \frac{|f(\alpha)|}{m}$, como queríamos demostrar. Como m puede tomarse, por ejemplo, el $\min_{x \in [a,b]} |f'(x)|$ siempre que este número exista y sea positivo.

En los apartados siguientes de este tema estudiaremos los métodos básicos de resolución de ecuaciones no lineales (bisección, Lagrange, secante, Newton-Raphson, etc.), algunos resultados para el caso particular de las ecuaciones polinomiales, un método de aceleración de la convergencia y una introducción a la resolución numérica de sistemas de ecuaciones no lineales.

2.2. Ecuaciones no lineales con una variable

2.2.1. El método de bisección

Es quizás el método más sencillo de aproximación de las raíces de una ecuación $f(x) = 0$, pero en general, bastante lento; aunque puede utilizarse como punto de partida para otro tipo de métodos.

Pasemos a exponerlo brevemente, en primer lugar sea una ecuación $f(x) = 0$, con f continua en $[a, b]$ y tal que $f(a)f(b) < 0$, es decir f posee una raíz en $[a, b]$, que en lo sucesivo supondremos única. Para buscar dicha raíz, dividimos el intervalo $[a, b]$ en dos partes iguales por su punto medio $\frac{a+b}{2}$; si $f(\frac{a+b}{2}) = 0$ entonces $r = \frac{a+b}{2}$ es la raíz buscada, en otro caso tomamos aquella de las dos mitades $[a, \frac{a+b}{2}]$ o $[\frac{a+b}{2}, b]$ en que la función toma signos opuestos en los extremos, el nuevo segmento que designamos por $[a_1, b_1]$ lo volvemos a dividir en dos partes iguales por su punto medio y reiteramos el mismo razonamiento anterior; así en una cierta etapa se obtiene la raíz o una sucesión de intervalos encajados

$$[a, b] \supset [a_1, b_1] \supset [a_2, b_2] \supset \cdots \supset [a_n, b_n] \supset \cdots$$

tales que $f(a_n)f(b_n) < 0 (n = 1, 2, ...)$ y $b_n - a_n = \frac{b-a}{2^n}$. Puesto que los extremos inferiores de dichos intervalos $\{a_n\}$ forman una sucesión creciente y acotada superiormente por b y los extremos superiores de los mismos $\{b_n\}$ forman una sucesión decreciente y acotada inferiormente por a, existen los límites

$$lim_{n\to\infty} a_n = \alpha \leq b$$
$$lim_{n\to\infty} b_n = \beta \geq a$$

como además $lim_{n\to\infty}(a_n - b_n) = \alpha - \beta = 0$ se sigue que $\alpha = \beta \in [a, b]$, siendo este límite común la raíz buscada, pues al ser $f(a_n)f(b_n) < 0$ por la continuidad de f se deduce que $lim_{n\to\infty} f(a_n)f(b_n) = f(\alpha)^2 \leq 0$, por tanto $f(\alpha) = 0$ y $\alpha = r$ es la raíz buscada de $f(x) = 0$ en el intervalo $[a, b]$.

Cota de error absoluto de la enésima aproximación. Si como valor de la raíz buscada r tomamos la aproximación enésima por defecto a_n, se tiene la siguiente acotación del error $0 \leq r - a_n \leq \frac{1}{2^n}(b - a)$; si se toma la aproximación por exceso b_n se tiene la acotación $0 \leq b_n - r \leq \frac{1}{2^n}(b - a)$, en tanto que si se toma como aproximación de la raíz r buscada el punto medio del enésimo intervalo $x_n = \frac{a_n+b_n}{2}$, entonces tenemos la siguiente acotación para el error $| r - x_n | \leq \frac{b-a}{2^{n+1}}$, es por tanto una mejor aproximación de r, pero ahora no podemos precisar su posición relativa a r, es decir si se trata de una aproximación por exceso o por defecto.

Ejercicio. Probar que la ecuación $x^3 - x - 1 = 0$ posee una única raíz en $[1, 2]$ y aproximarla por el método de bisección con error menor que $5 \cdot 10^{-6}$.

Solución. Sea $f(x) = x^3 - x - 1$ es fácil ver que $f(1) = -1 < 0$ y $f(2) = 5 > 0$, luego existe, al menos, una raíz $s \in [a, b]$, ahora bien $f'(x) = 3x^2 - 1$ sólo se anula en los puntos $\frac{1}{\sqrt{3}}$ y $-\frac{1}{\sqrt{3}}$, ambos fuera del intervalo $[1, 2]$, por tanto $f'(x)$ tiene signo constante en dicho intervalo, en este caso es $f'(x) > 0$, por tanto la función $f(x)$ es estrictamente creciente, en consecuencia la raíz r de $f(x) = 0$ en $[1, 2]$ es única. Para aproximarla por el método de bisección con error menor que $5 \cdot 10^{-6}$, basta con tomar n tal que $\frac{1}{2^n} < 5 \cdot 10^{-6}$ o sea $n > \log(200000)/log(2) = 17{,}60964047443681$, es decir con 18 pasos por el método de bipartición nos aseguraremos la aproximación deseada tomando cualquier punto de ese subintervalo (por lo comentado antes, si tomamos el punto medio basta con 17 pasos). Dando el punto medio del subintervalo obtenido para $n = 18$, como aproximación de la raíz, resulta $s \simeq 1{,}324716567993164$, en doble precisión con Maxima obtenemos el valor $s = 1{,}324717957244746$.

2.2.2. Teorema del punto fijo: método iterativo general

Dada una ecuación $f(x) = 0$ con f continua y una única raíz r en el intervalo cerrado $[a, b]$, se quiere aproximar dicha raíz, para lo cual se pasa de alguna manera a una ecuación equivalente de la forma $x = g(x)$ (es decir con las mismas raíces, en este caso con la misma única raíz r en dicho intervalo); entonces $f(r) = 0$ si y sólo si $r = g(r)$, o sea r es raíz de f si y sólo si r es un punto fijo de la aplicación g. Para aproximar esta raíz se parte de algún punto $x_0 \in [a, b]$ se genera la sucesión $\{x_n\}$ en la forma $x_{n+1} = g(x_n)$, el problema que se plantea es saber elegir un método, es decir una función g de modo que la sucesión $\{x_n\}$ converja a la raíz buscada r. El siguiente **teorema de punto fijo** nos da condiciones suficientes para la existencia y unicidad de dicha raíz, así como un método que genera una sucesión convergente a la misma.

Teorema 4 ***(Teorema del punto fijo)*** *Si g es una función real definida en el intervalo $[a, b]$, que cumple las dos condiciones siguientes:*

1) $\forall x \in [a, b] \Rightarrow g(x) \in [a, b]$

2) Existe un número real k verificando $0 \leq k < 1$ tal que $\forall x, y \in [a, b] \Rightarrow | g(x) - g(y) | \leq k \mid x - y \mid$

Entonces, existe una única raíz r de la ecuación $x = g(x)$ en $[a, b]$, que se obtiene como límite de una sucesión $\{x_n\}$ donde x_0 es un punto arbitrario de $[a, b]$ y $x_{n+1} = g(x_n)$ para $n \geq 0$.

Demostración. La demostración directa es sencilla, pero en realidad este teorema es caso particular del **teorema de la aplicación contractiva** (cuya demostración no haremos) que afirma que si E es un espacio métrico completo (es decir un espacio métrico en el que toda sucesión de Cauchy en E es convergente), $g : E \to E$ una aplicación tal que existe un número real k verificando $0 \leq k < 1$ y que $\forall x, y \in E$ es $d(g(x), g(y)) \leq kd(x, y)$ (una tal aplicación se denomina una **contracción** de E), entonces existe una única raíz r de la ecuación $x = g(x)$ en E, que se obtiene como límite de una sucesión $\{x_n\}$, donde x_0 es un punto cualquiera de E y $x_{n+1} = g(x_n)$ para $n \geq 0$.

Ya que si g verifica las condiciones (1) y (2) anteriores es una contracción en $[a, b]$, puesto que este intervalo es un cerrado en $\mathbb{R}$ es por tanto un espacio métrico completo con la $d(x, y) =\mid x - y \mid$, y en virtud del teorema citado se sigue la tesis.

Observaciones:

1. Cuando una función verifica la condición (2) del teorema anterior con k cualquiera ($k \geq 0$) se dice que es lipschitziana en $[a, b]$ o que verifica una condición de Lipschitz.

2. Si g es derivable con derivada continua y cumple la condición:

$$\mid g'(x) \mid \leq k < 1 \quad \forall x \in [a, b] \quad (2')$$

   entonces, por el teorema del valor medio, g verifica la condición (2) del teorema anterior. Puede tomarse como $k = \text{máx}_{x\in[a,b]} |g'(x)|$ siempre que este k sea menor que 1.

3. El teorema anterior con la condición (2) o la (2′) sigue siendo válido si se sustituye el intervalo $[a, b]$ por cualquier otro intervalo cerrado de $\mathbb{R}$ no necesariamente acotado, I, pues este constituiría un espacio métrico completo, ya que todo subconjunto cerrado de un espacio métrico completo es el mismo un espacio métrico completo.

4. Puede probarse sin dificultad que los errores en el paso enésimo verifican las acotaciones siguientes:

$$\boxed{\mid x_n - r \mid \leq \frac{k}{1-k} \mid x_n - x_{n-1} \mid \leq \frac{k^n}{1-k} \mid x_1 - x_0 \mid}$$

la primera nos sirve para estimar el error a posteriori y la última para estimarlo a priori y determinar el número de iteraciones a realizar para garantizar una aproximación requerida.

5. El teorema anterior (o sus variantes apuntadas) nos dan resultados de convergencia **global** pues la aseguran para todo punto inicial x_0 en el intervalo dado, pero sus condiciones son muy restrictivas, por ello estamos interesados en resultados de convergencia **local** como el que sigue.

Teorema 5 ***(Condición suficiente de convergencia local)*** *Sea g de clase $\mathbb{C}^{(1)}$ en algún intervalo que contenga a un punto fijo r de g. Entonces, si $\mid g'(r) \mid < 1$, existe $\delta > 0$ tal que $\forall x_0 \in [r-\delta, r+\delta]$ se tiene que la sucesión $\{x_n\}_0^\infty$ con $x_{n+1} = g(x_n)(n \geq 0)$ converge a r.*

Demostración. Si $\mid g'(r) \mid < 1$ por continuidad de la derivada $\exists k < 1$ y $\delta > 0$ tal que $\mid g'(x) \mid \leq k$, en el entorno $[r - \delta, r + \delta]$. Ahora bien, en dicho entorno se cumple la condición (2) del teorema 4 anterior. Para ver que también se cumple la condición (1), basta con tener en cuenta que si $\mid x - r \mid \leq \delta$ se sigue que $\mid g(x) - r \mid = \mid g(x) - g(r) \mid \leq k \mid x - r \mid < \delta$ lo cual implica que $g(x) \in [r - \delta, r + \delta]$ como queríamos demostrar. Este teorema nos garantiza que en las condiciones del mismo, si elegimos un punto inicial suficientemente próximo a la raíz buscada el algoritmo converge a dicha raíz, esto es lo mínimo que exigiremos a cualquier método iterativo de este tipo.

Orden de convergencia de un método

Además de la convergencia de un método dado, se quiere saber si la sucesión definida por las iteraciones $x_{n+1} = g(x_n)$ converge rápidamente a la raíz buscada r; es decir como disminuye el error $e_n = x_n - r$ de una iteración a la siguiente, lo que nos conducirá a la siguiente definición de orden de un método iterativo.

Definición 1 *Se dice que el método $x_{n+1} = g(x_n)$ es **convergente de orden** 1 o que tiene **convergencia lineal** si $lim_{n\to\infty}\frac{|e_{n+1}|}{|e_n|} = k$ (siempre que sea $0 < k < 1$). Y se dice **convergente de orden** p, con $p > 1$, si $lim_{n\to\infty}\frac{|e_{n+1}|}{|e_n|^p} = k > 0$ (en este caso basta con que k sea positivo).*

Para estudiar el orden de un método, definido por la función de iteración g, si esta es suficientemente diferenciable, se recurre al desarrollo de Taylor, de modo que si $g \in \mathbb{C}^{(k+1)}(\mathbb{I})$ en un intervalo $\mathbb{I}$ que contiene a r se tendrá:

$$e_{n+1} = x_{n+1} - r = g(x_n) - g(r) =$$
$$g'(r)e_n + \frac{1}{2}g''(r)e_n^2 + ... + \frac{g^{(k)}(r)}{k!}e_n^k + \frac{g^{(k+1)}(\xi_n)}{(k+1)!}e_n^{k+1}$$

Por tanto $lim_{n\to\infty}\frac{|e_{n+1}|}{|e_n|} =| g'(r) |$, luego si $0 <| g'(r) |< 1$ se dice que el método tiene **convergencia lineal o de primer orden**. Si $g'(r) = 0$, se deduce que $lim_{n\to\infty}\frac{|e_{n+1}|}{|e_n|^2} = \frac{1}{2} | g''(r) |$, luego si $g''(r) \neq 0$ se dice que la **convergencia es cuadrática o de segundo orden**. Y, en general, si se verifican las condiciones $g'(r) = g''(r) = ... = g^{(p-1)}(r) = 0$ y $g^{(p)}(r) \neq 0$ el método será de **orden p**. En general, si se verifican las igualdades $g'(r) = g''(r) = ... = g^{(p-1)}(r) = 0$, pero no sabemos si $g^{(p)}(r)$ se anula o no, decimos que la convergencia es **de orden, al menos,** p. Para otros métodos cuyo error no se comporta de esta forma el orden de convergencia puede no ser entero y se estudia por otros procedimientos.

2.2.3. Métodos iterativos particulares de aproximación de soluciones

En cada uno de los métodos que se exponen a continuación, se parte de una ecuación $f(x) = 0$, con f real continua en $[a, b]$ y tal que $f(a)f(b) < 0$, es decir f posee una raíz r en $[a, b]$, que supondremos única.

Método de aproximaciones sucesivas

Aunque los métodos definidos por el teorema 4 anterior son todos métodos de aproximaciones sucesivas, se suele conocer como tal al método que consiste en pasar de la ecuación $f(x) = 0$ a la ecuación equivalente $x = x - f(x) = g(x)$. Ahora bien si f y f' son continuas también lo serán g y g', y para tener convergencia, al menos, local del método iterativo

$$x_0 \text{ dado}$$
$$x_{n+1} = x_n - f(x_n)$$

es suficiente, de acuerdo con el teorema de convergencia local (teorema 5 anterior), que si r es la raíz buscada se tenga $| g'(r) |=| 1 - f'(r) |< 1$ o equivalentemente $0 < f'(r) < 2$. Puesto que esta condición es muy restrictiva, a veces suele introducirse una función real no nula $\alpha(x)$ de la manera siguiente

$$x_{n+1} = x_n - \frac{f(x_n)}{\alpha(x_n)}$$

En particular, si $\alpha(x_n) = \alpha$ es constante se tiene que $g'(r) = 1 - \frac{f'(r)}{\alpha}$ con lo cual basta con elegir α para que $| 1 - \frac{f'(r)}{\alpha} |< 1$; además, de acuerdo con el apartado anterior si $g'(r) = 1 - \frac{f'(r)}{\alpha} \neq 0$ el método correspondiente tendrá convergencia lineal o de orden uno.

Método de Lagrange

El método de Lagrange, también conocido "como método de las partes proporcionales o regula falsi", consiste básicamente en reemplazar la gráfica de f restringida al intervalo $[a, b]$ por la recta pasando por los puntos extremos $A(a, f(a))$ y $B(b, f(b))$, es decir se sustituye la función f por un polinomio de grado uno $p(x)$ y se resuelve la ecuación $p(x) = 0$; el punto de intersección de la recta AB con el eje de las x está dado por

$$x_1 = a - f(a)\tfrac{b-a}{f(b)-f(a)} = \tfrac{af(b)-bf(a)}{f(b)-f(a)}$$

Seguidamente se calcula $f(x_1)$ y se determina en cual de los dos intervalos $[a, x_1]$ o $[x_1, b]$ está la raíz buscada r, suponiendo que $r \in [a, x_1]$, entonces se puede volver a aplicar el mismo procedimiento cambiando b por x_1, obteniéndose

$$x_2 = a - f(a)\tfrac{x_1-a}{f(x_1)-f(a)} = \tfrac{af(x_1)-x_1f(a)}{f(x_1)-f(a)}$$

En general, si esto es posible en cada paso, es decir el extremo a permanece fijo y cambia el b, se tendrá el siguiente algoritmo

$$\boxed{x_0 = b,\ x_{n+1} = g(x_n)\ (n \geq 0)}$$

con g definida por la expresión

$$\boxed{g(x) = \frac{af(x) - xf(a)}{f(x) - f(a)}}$$

Con objeto de dar condiciones suficientes de convergencia local, supongamos que $f(x)$ es derivable, entonces para el algoritmo obtenido será $g'(r) = \frac{f(a)+(r-a)f'(r)}{f(a)}$ y si f es de clase $\mathbb{C}^{(2)}$ en un intervalo que contenga al $[a, r]$ entonces desarrollando por Taylor se tiene $f(a) = f'(r)(a-r) + \frac{f''(c)}{2}(a-r)^2$ con $a < c < r$ (pues $f(r) = 0$), luego una condición suficiente para que haya convergencia local, es que se verifique $\mid g'(r) \mid = \mid \frac{f''(c)(a-r)^2}{2f(a)} \mid < 1$, y la convergencia será de primer orden o lineal si además es $g'(r) \neq 0$.

Obsevación. En el supuesto anterior, el punto inicial es el extremo b, pero podría darse el caso de que el punto inicial fuera el extremo a. En particular, se puede probar que si f es de clase $\mathbb{C}^{(2)}$ en el intervalo $[a, b]$, $f(a)f(b) < 0$, $f'(x) \neq 0$ y $f''(x) \neq 0$ en dicho intervalo, entonces existe una única raíz y además el método converge dejando fijo un extremo, aquel en el que la función y la derivada segunda tienen el mismo signo y tomando el otro como valor inicial del algoritmo de iteración. Los algoritmos correspondientes

para todos los casos posibles, en estas condiciones, se reducen al anterior y al siguiente:

$$\boxed{x_0 = a,\ x_{n+1} = g(x_n)\ (n \geq 0)}$$

con g definida por la expresión

$$\boxed{g(x) = \frac{xf(b) - bf(x)}{f(b) - f(x)}}$$

Método de Newton-Raphson

Si $f(x) = 0$ posee una raíz en el punto r del intervalo $[a, b]$, la idea de este método consiste en reemplazar la curva $f(x)$ por la recta tangente a la misma en uno de sus puntos extremos. Suponer, de momento, que se toma la tangente en el punto b, cuya ecuación es $y - f(b) = f'(b)(x - b)$ y su intersección con el eje OX está dada por el punto de abscisa

$$x_1 = b - \frac{f(b)}{f'(b)}$$

que representa un valor aproximado de r. Si ahora volvemos a trazar la tangente por el punto de abscisa x_1 y cortamos con el eje OX se obtiene la segunda aproximación de la raíz en la forma

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Reiterando este modo de proceder, obtenida la aproximación n-sima se obtiene la (n+1)-sima mediante el algoritmo siguiente

$$\boxed{x_0 = b\ ,\ x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\quad (n \geq 0)}$$

o también

$$\boxed{x_0 = b\ ,\ x_{n+1} = g(x_n)\ (n \geq 0)}$$

con

$$\boxed{g(x) = x - \frac{f(x)}{f'(x)}}$$

El método estará bien definido para x_n próximo a r con la condición $f'(r) \neq 0$; es decir r es un cero simple de f, pues si $f'(r) \neq 0$, siendo f' continua y x_n suficientemente próximo a r se tendrá $f'(x_n) \neq 0$.

Otra forma de deducir este método es la siguiente; dada la ecuación $f(x) = 0$ pasemos a la ecuación equivalente

$$x = x - \frac{f(x)}{\alpha(x)} = g(x)$$

con $\alpha(x) \neq 0$; ahora si dichas funciones son derivables, se tendrá

$$g'(x) = 1 - \frac{f'(x)\alpha(x) - \alpha'(x)f(x)}{\alpha(x)^2}$$

y si r es una raíz de $f(x) = 0$ obtendremos $g'(r) = 1 - \frac{f'(r)}{\alpha(r)}$, por tanto siempre que $\alpha(x)$ verifique $\alpha(r) = f'(r)$ la convergencia del método $x_{n+1} = g(x_n)$ será, al menos, cuadrática, por lo que basta con tomar $\alpha(x) = f'(x)$, si $f'(x)$, no se anula para que el método

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \ (n \geq 0)$$

tenga convergencia, al menos, cuadrática.

Seguidamente, damos sendos resultados relativos a la convergencia local y global del método de Newton (también llamado de Newton-Raphson).

Teorema 6 ***(Teorema de convergencia local)*** *Sea f'' continua y f' no nula en algún intervalo abierto que contenga la raíz r de $f(x) = 0$. Entonces, existe $\varepsilon > 0$ tal que el método de Newton es convergente para todo x_0 tal que $| x_0 - r | \leq \varepsilon$. Además, si f''' es continua la convergencia es, al menos, cuadrática.*

Teorema 7 ***(Teorema de convergencia global-1)*** *Sea $f \in \mathbb{C}^{(2)}([a, b])$ verificando:*

1. *$f(a) \cdot f(b) < 0$*

2. *$\forall x \in [a, b]$ es $f'(x) \neq 0$ (monotonía estricta)*

3. *$\forall x \in [a, b]$ es $f''(x) \geq 0$ (o $\forall x \in [a, b]$ es $f''(x) \leq 0$, concavidad en el mismo sentido)*

Entonces, existe una única raíz r de $f(x) = 0$ en $[a, b]$ y la sucesión $\{x_n\}_0^\infty$, definida por el algoritmo del método de Newton converge, hacia r para todo $x_0 \in [a, b]$ tal que $f(x_0) \cdot f''(x_0) \geq 0$. Si además $f \in \mathbb{C}^{(3)}([a, b])$ la convergencia es, al menos, cuadrática.

Teorema 8 ***(Teorema de convergencia global-2)*** *Sea $f \in \mathbb{C}^{(2)}([a, b])$ verificando:*

1. *$f(a) \cdot f(b) < 0$*

2. *$\forall x \in [a, b]$ es $f'(x) \neq 0$ (monotonía estricta)*

3. $\forall x \in [a,b]$ *es* $f''(x) \geq 0$ *(o* $\forall x \in [a,b]$ *es* $f''(x) \leq 0$*, concavidad en el mismo sentido)*

4. $\max \mid \frac{f(a)}{f'(a)} \mid, \mid \frac{f(b)}{f'(b)} \mid \leq b - a$

Entonces, existe una única raíz r de $f(x) = 0$ *en* $[a,b]$ *y la sucesión* $\{x_n\}_0^\infty$*, definida por el algoritmo del método de Newton converge, hacia r para todo* $x_0 \in [a,b]$*. Si además* $f \in \mathbb{C}^{(3)}([a,b])$ *la convergencia es, al menos, cuadrática.*

Método de la secante

El método de Newton es, en general, de convergencia cuadrática, obteniendo buenas aproximaciones en pocas iteraciones, pero tienen el inconveniente de hacer intervenir la derivada $f'(x)$ de la función $f(x)$, y puede suceder que dicha derivada sea difícil de calcular, cabe entonces la posibilidad de aproximar $f'(x_n)$ en el método de Newton por un cociente en diferencias de la forma

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

dando lugar al método de la secante, cuyo algoritmo adopta la forma

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} \ (n \geq 1)$$

En este caso x_{n+1} se obtiene en función de x_n y x_{n-1} y no sólo en función de x_n como ocurría en los métodos precedentes. Si se compara con el método de Lagrange definido por el algoritmo

$$x_{n+1} = \frac{a f(x_n) - x_n f(a)}{f(x_n) - f(a)}$$

se observa que el método de la secante se obtiene reemplazando a por x_{n-1}, de manera que x_{n+1} es la abscisa de la intersección con el eje Ox de la recta que pasa por los dos puntos $(x_n, f(x_n))$ y $(x_{n-1}, f(x_{n-1}))$. Cabe esperar que este método tenga un orden superior al de Lagrange e inferior al de Newton, aunque dado que no es un método de la forma $x_{n+1} = g(x_n)$ el estudio de su orden de convergencia se hace de modo diferente, se tiene al respecto el siguiente resultado.

Teorema 9 *El orden de convergencia del método de la secante es* $p = \frac{1+\sqrt{5}}{2} = 1{,}618\ldots$.

El método necesita para iniciarse dos valores x_0 y x_1, que pueden ser los extremos del intervalo que contiene a la raíz buscada, o bien partiendo de un x_0 inicial, calcular x_1 por algún otro método (por ejemplo aplicando el método de Newton).

2.2.4. Aceleración de la convergencia. El método $\triangle^2$ de Aitken

Para el método de aproximaciones sucesivas y el de Lagrange, así como para cualquier otro método de orden uno, los errores de las aproximaciones sucesivas de la raíz buscada verifican

$$e_{n+1} = (A + \varepsilon_n)e_n \text{ con } A = g'(r),\ 0 <\mid A \mid< 1 \text{ y } lim_{n\to\infty}\varepsilon_n = 0$$

Si se supone que $\varepsilon_n = 0$, entonces se tiene

$$\begin{aligned} x_{n+1} - r &= A(x_n - r) \\ x_{n+2} - r &= A(x_{n+1} - r) \end{aligned}$$

por tanto, restando a la segunda la primera, deducimos $x_{n+2} - x_{n+1} = A(x_{n+1} - x_n)$, de donde resulta

$$A = \frac{x_{n+2}-x_{n+1}}{x_{n+1}-x_n}$$

luego de la primera de las ecuaciones anteriores se deduce

$$r = \frac{1}{1-A}(x_{n+1} - Ax_n) = x_n + \frac{1}{1-A}(x_{n+1} - x_n)$$

y sustituyendo A por su valor, finalmente, se obtiene

$$r = x_n - \frac{(x_{n+1}-x_n)^2}{x_{n+2}-2x_{n+1}+x_n}$$

Es decir bajo la hipótesis de que el método sea de orden uno y con $\varepsilon_n = 0$, se obtiene la solución exacta con sólo tres iteraciones sucesivas. Ahora bien, en el caso de un método de primer orden para el que $\varepsilon_n \neq 0$, pero sea pequeño en comparación con A en el entorno del límite, se puede prever que la sucesión $\{x'_n\}$ definida por

$$\boxed{x'_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n} = x_n - \frac{(\triangle x_n)^2}{\triangle^2 x_n}}$$

sea una mejor aproximación de r que x_n, y eso es precisamente lo que afirma el teorema siguiente, que damos sin demostración. Recordemos que el operador $\triangle$ actúa sobre sucesiones produciendo nuevas sucesiones en la forma $\triangle x_n = x_{n+1} - x_n$, en tanto que $\triangle^2$ actúa en la forma $\triangle^2 x_n = \triangle(\triangle x_n) = \triangle x_{n+1} - \triangle x_n = x_{n+2} - 2x_{n+1} + x_n$.

Teorema 10 *Si $\{x_n\}$ es una sucesión que converge hacia r, con convergencia de orden uno, entonces la sucesión $\{x'_n\}$ definida anteriormente converge a r más rápidamente, es decir que*

$$lim_{n\to\infty}\frac{x'_n - r}{x_n - r} = 0$$

Se conoce como método de aceleración de la convergencia $\triangle^2$ de Aitken.

Ejercicio. Aproximar la única raíz de la ecuación $f(x) = x^3 - x - 1 = 0$ en el intervalo $[1, 2]$ con error menor que $5 \cdot 10^{-6}$ por los métodos de bipartición, Lagrange, Lagrange acelerado y Newton.

2.3. Ecuaciones polinomiales

Los métodos anteriormente expuestos son aplicables a la resolución de ecuaciones polinómicas $p(x) = 0$, siendo

$$p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0 \text{ con } a_i \in \mathbb{R} \text{ y } a_0 \neq 0$$

pero hay diversas peculiaridades relativas a la evaluación de polinomios y sus derivadas, así como a la acotación de sus raíces, que por su eficiencia pasamos a relatar. En particular, para evaluar un polinomio utilizaremos el algoritmo de Horner, visto en la lección anterior, que requiere tan sólo un total de $2n$ operaciones (n sumas y n multiplicaciones).

2.3.1. Acotación de raíces de una ecuación polinómica

En general, dada una ecuación $f(x) = 0$, se dice que el número real L es una **cota superior** de sus raíces reales si para toda raíz real r de dicha ecuación se verifica que $r \leq L$. Análogamente se dice que el número real l es una **cota inferior** de sus raíces si verifica que $l \leq r$. Y se dice que se han **acotado** sus raíces reales si se ha determinado un intervalo $[l, L]$ que las contiene a todas. Si se admite la posibilidad de raíces complejas, se dirá que se han **acotado** si existe $M \geq 0$ tal que $\mid r \mid \leq M$ para toda raíz r de $f(x) = 0$. Antes de proseguir, hagamos la siguiente observación.

Observación. Notemos que r es solución de la ecuación $f(x) = 0$ si y sólo si $-r$ lo es de $f(-x) = 0$ (o también de la ecuación equivalente a esta última $-f(-x) = 0$). Asimismo, r es una raíz no nula de $f(x) = 0$ si y sólo si $1/r$ lo es de $f(1/x) = 0$ (o también de la ecuación equivalente, para hallar raíces no nulas, $x^n f(1/x) = 0$).

Ahora, veamos algunos resultados sencillos de acotación de raíces de ecuaciones polinómicas.

Teorema 11 *Sea la ecuación polinómica* $p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$ *con* $a_0 \neq 0$*, entonces si* $\lambda = \max\{\mid \frac{a_i}{a_0} \mid\mid 1 \leq i \leq n\}$*, toda raíz de* $p(x) = 0$ *verifica que* $\mid r \mid \leq M = \lambda + 1$*. Es decir todas las raíces de la ecuación están contenidas en el círculo de radio* $\lambda + 1$*. Además, si* $a_n \neq 0$ *entonces toda raíz* r *es no nula y verifica que* $\mid r \mid \geq m = \frac{1}{\lambda'+1}$*, siendo* $\lambda' = \max\{\mid \frac{a_i}{a_n} \mid\mid 0 \leq i \leq n-1\}$*, por tanto si tanto* a_0 *como* a_n *son no nulas todas las raíces de* $p(x) = 0$ *están en la corona de centro el origen y radios* m *y* M*.*

Demostración. Si r es una raíz de dicha ecuación, entonces se tiene $a_0r^n + a_1r^{n-1} + \cdots + a_{n-1}r + a_n = 0$ de donde se deduce que

$$\mid r \mid^n = \mid \frac{a_1}{a_0}r^{n-1} + \cdots + \frac{a_{n-1}}{a_0}r + \frac{a_n}{a_0} \mid \leq \lambda(\mid r \mid^{n-1} + \cdots + \mid r \mid + 1) = \lambda\frac{|r|^n - 1}{|r| - 1}$$

Si siendo r raíz fuese $\mid r \mid > \lambda + 1$ se tendría que $\mid r \mid -1 > \lambda$ o equivalentemente $\frac{1}{|r|-1} < \frac{1}{\lambda}$ y entonces cumpliría $\mid r \mid^n < \lambda\frac{|r|^n - 1}{\lambda} = \mid r \mid^n - 1$ lo cual es absurdo, por tanto para toda raíz r se debe verificar que $\mid r \mid \leq \lambda + 1$, como queríamos demostrar. Para la segunda parte basta cambiar x por $1/x$ en la ecuación $p(x) = 0$.

Ejercicio. Acotar las raíces de la ecuación $f(x) = x^3 - x - 1 = 0$.

2.3.2. Determinación de cotas superiores de las raíces reales de una ecuación polinómica

Veamos brevemente tres métodos para obtener una cota superior L de las raíces reales de una ecuación polinómica de grado $n \geq 1$. Puesto que las ecuaciones $f(x) = 0$ y $f(-t) = 0$ tienen raíces opuestas, si L' es una cota superior de las raíces reales de $f(-t) = 0$ entonces $l = -L'$ es una cota inferior de las raíces reales de $f(x) = 0$, en consecuencia si sabemos hallar una cota superior también sabemos obtener una cota inferior.

1. **(Laguerre-Thibault)** Si $L \geq 0$ y en la división de $p(x)$ por $x - L$ son no negativos todos los coeficientes del cociente y también el resto, entonces L es una cota superior de las raíces reales de $p(x)$.

   Demostración. En efecto, $p(x) = p(L) + (x-L)(b_0x^{n-1} + \cdots + b_{n-2}x + b_{n-1})$ y si $p(L) \geq 0$ y $b_i \geq 0$ $(i = 0, 1, \ldots, n-1)$, entonces para todo $x > L \geq 0$ real resulta ser $p(x) > 0$ y x no puede ser raíz, de donde se sigue el resultado.

2. **Newton** Si $p(L) \geq 0, p'(L) \geq 0, \ldots, p^{(n)}(L) \geq 0$ entonces L es una cota superior de las raíces reales de $p(x) = 0$.

Demostración. Puesto que $p(x) = p(L)+p'(L)(x-L)+\frac{p''(L)}{2!}(x-L)^2+\cdots+\frac{p^{(n)}(L)}{n!}(x-L)^n$, entonces para todo $x > L$ real resulta $p(x) > 0$, luego L es una cota superior de las raíces reales.

3. **Método de agrupación de términos.** Supongamos la ecuación ordenada y con el primer coeficiente a_0 positivo (si fuera necesario multiplicamos por -1), seguidamente agrupemos los términos de la misma de manera que cada grupo empiece por un término con coeficiente positivo y no presente más que una variación de signo, supongamos que para un cierto $\xi_0 > 0$ el primero de dichos grupos es mayor o igual que cero, entonces también lo es para todo $x > \xi_0$, si designamos por L al mayor de los valores ξ_i que corresponden a cada grupo, entonces todos los grupos son positivos para $x > L$, o sea no hay raíces reales mayores que L, y por tanto L es una cota superior de sus raíces reales. Para hallar una cota inferior de las raíces reales, basta con hallar por este mismo procedimiento una cota superior L' de las raíces reales de la ecuación $p(-x) = 0$, cuyas raíces son opuestas a las de $p(x) = 0$ y tomar $l = -L'$.

Ejercicio. Dada la ecuación $p(x) = x^7-2x^6+x^5-x^4+2x^3-3x^2+7x-1 = 0$, se pide acotar sus raíces y obtener (utilizando alguno de los métodos anteriores) cotas superiores e inferiores de sus raíces reales.

2.4. Sistemas no lineales. Método de Newton para sistemas

Vamos a restringirnos a sistemas de dos ecuaciones con dos incógnitas, aunque para el caso general de n ecuaciones con n incógnitas estos métodos se generalizan sin dificultad alguna.

2.4.1. Método de iteración simple en varias variables

Para fijar ideas sea un sistema de dos ecuaciones con dos incógnitas

$$\begin{aligned} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{aligned}$$

lo escribimos en forma equivalente como

$$\begin{aligned} x_1 = g_1(x_1, x_2) \\ x_2 = g_2(x_1, x_2) \end{aligned}$$

con lo cual el problema se traduce en buscar un punto fijo de la función vectorial de dos variables $g(x) = g(x_1, x_2) = (g_1(x_1, x_2), g_2(x_1, x_2))$.

Teorema 12 *Sea C un cerrado de $\mathbb{R}^2$ tal que se verifican las condiciones:*

I) $\forall x \in C \Rightarrow g(x) \in C$

II) $\exists L$ *con* $0 \leq L < 1$*tal que* $\forall x, y \in C \Rightarrow \|g(x) - g(y)\| \leq L \parallel x - y \parallel$

Entonces, para todo $x^{(0)} = (x_1^{(0)}, x_2^{(0)}) \in C$, la sucesión $\{x^{(m)}\}_0^\infty = \{(x_1^{(m)}, x_2^{(m)})\}_0^\infty$ definida por $x^{(m+1)} = g(x^{(m)})$ o también

$$\begin{aligned} x_1^{(m+1)} &= g_1(x_1^{(m)}, x_2^{(m)}) \\ x_2^{(m+1)} &= g_2(x_1^{(m)}, x_2^{(m)}) \end{aligned}$$

converge a la única solución $r = (r_1, r_2))$ del sistema de ecuaciones

$$\begin{aligned} x_1 &= g_1(x_1, x_2) \\ x_2 &= g_2(x_1, x_2) \end{aligned}$$

Además, para todo $m \geq 1$ se tiene $\|x^{(m)} - r\| \leq \frac{L^m}{1-L} \parallel x^{(1)} - x^{(0)} \parallel$.

Demostración. Es una consecuencia del teorema de la aplicación contractiva de un espacio métrico completo.

2.4.2. El método de Newton para sistemas

Dado el sistema

$$\begin{aligned} f_1(x_1, x_2) &= 0 \\ f_2(x_1, x_2) &= 0 \end{aligned}$$

con $f(x_1, x_2) = (f_1(x_1, x_2), f_2(x_1, x_2))$ de clase $\mathbb{C}^{(3)}$ en un entorno de la raíz buscada $r = (r_1, r_2)$ y con jacobiano inversible en r, pasamos al sistema equivalente

$$\begin{aligned} x_1 &= g_1(x_1, x_2) \\ x_2 &= g_2(x_1, x_2) \end{aligned}$$

donde ahora es $g(x) = x - J_f^{-1}(x) f(x)$. El método definido por esta g se denomina método de **Newton para sistemas**, es localmente convergente con convergencia de segundo orden y viene dado por el algoritmo

$$\boxed{x^{(m+1)} = x^{(m)} - J_f^{-1}(x^{(m)}) f(x^{(m)})}$$

La aplicación de este método requiere que $J_f(x^{(m)})$ sea inversible para todo m y en la práctica se suele presentar en la forma

$$J_f(x^{(m)})(x^{(m+1)} - x^{(m)}) = -f(x^{(m)})$$

y llamando $\delta^{(m)} = x^{(m+1)} - x^{(m)}$, el método consiste en

$$\text{Hallar } \delta^{(m)} \text{ verificando } J_f(x^{(m)})\delta^{(m)} = -f(x^{(m)})$$
$$\text{y obtener } x^{(m+1)} = x^{(m)} + \delta^{(m)}$$

Ejercicio. Hacer un par de iteraciones por el método de Newton para aproximar la raíz del sistema

$$x^3 - 3xy^2 + 1 = 0$$
$$3x^2y - y^3 = 0$$

partiendo del punto inicial $(x_0, y_0) = (1, 1)$ (la solución exacta es $x = \frac{1}{2}$ e $y = \frac{\sqrt{3}}{2} \cong 0{,}866025$).

2.5. Problemas resueltos

1. Probar que la ecuación $e^x - 3x = 0$ posee una única raíz en el intervalo $[0, 1]$, realizar seis iteraciones por el método de bisección para encontrar la raíz aproximada. ¿Cuántos decimales correctos tiene dicha aproximación? ¿Cuántas iteraciones son necesarias para que la raíz obtenida tenga cuatro decimales significativos correctos?

   Solución. Sea $f(x) = e^x - 3x$, entonces $f(0) = 1 > 0$ y $f(1) = e - 3 < 0$, lo que asegura al ser $f(x)$ continua en el intervalo $[0, 1]$, que existe, al menos, una raíz en dicho intervalo, por otro lado dicha raíz es única pues la derivada $f'(x) = e^x - 3 < 0$ en dicho intervalo, ya que $f''(x) = e^x > 0$ por tanto $f'(x)$ es estrictamente creciente en $[0, 1]$ y siendo $f'(1) = -0{,}28171817154095 < 0$ es $f'(x) < 0$ en [0,1], por lo que f es estrictamente decreciente en $[0, 1]$, lo que asegura la unicidad de dicha raíz. Y realizando las seis iteraciones que nos piden se obtienen los siguientes resultados:

Extr. izquierdo	Extr. derecho	Amplitud intervalo	Pto. medio
0	1	1	$0,5\,(+)$
$0,5$	1	$0,5$	$0,75\,(-)$
$0,5$	$0,75$	$0,25$	$0,625\,(-)$
$0,5$	$0,625$	$0,125$	$0,5625\,(+)$
$0,5625$	$0,625$	$0,0625$	$0,59375\,(+)$
$0,59375$	$0,625$	$0,03125$	$0,609375$

Luego

$$\overline{x} = x_6 = 0,609375$$

Como

$$\frac{1}{2^n} \cdot (b-a) = \frac{1}{2^6} = 0,015 = 1{,}5 \cdot 10^{-2} < 5 \cdot 10^{-2}$$

Entonces la aproximación tendrá un decimal correcto, es decir, el 6.

Para calcular cuántas iteraciones son necesarias para obtener cuatro decimales significativos correctos, dando como aproximación el punto medio del $(n-1)$-simo intervalo, utilicemos la fórmula que liga el error asociado al número de subdivisiones. Así:

$$\frac{1}{2^n} \cdot (1-0) < 5 \cdot 10^{-5} \Leftrightarrow 2^n > 20{,}000 \Leftrightarrow n > \tfrac{\log(20000)}{\log(2)} = 14{,}2877...$$

Así pues, es suficiente con tomar $n = 15$ iteraciones para obtener cuatro decimales significativos correctos, lo que muestra la lentitud del método empleado.

2. Utilizar el método de bisección para encontrar una solución aproximada con error menor que 10^{-2} en el intervalo $[4, 4{,}5]$, para la ecuación:

$$x = \tan x$$

Solución. Como se desea una aproximación con error menor que 10^{-2}, se tendrá, dando el punto medio del último subintervalo, habrá de verificarse que:

$$\frac{1}{2^n} \cdot (4{,}5-4) < 10^{-2} \Leftrightarrow 2^n > 50 \Leftrightarrow n > \frac{\log(50)}{\log(2)} = 5{,}643856...$$

Por tanto, es suficiente con $n = 6$ iteraciones.

Así, $f(x) = x - \tan x$, $f(4) = 2{,}8421... > 0$ y $f(4{,}5) = -0{,}1373... < 0$.

El proceso iterativo resulta:

Extr. izquierdo	Extr. derecho	Pto. medio
4	4,5	4,25 (+)
4,25	4,5	4,375 (+)
4,375	4,5	4,3755 (+)
4,3755	4,5	4,46875 (+)
4,46875	4,5	4,484375 (+)
4,484375	4,5	4,4921875

Luego

$$\overline{x} = x_6 = 4{,}4921875$$

3. Sabiendo que existe una raíz de la ecuación:

$$x^3 + x = 6$$

entre $1,55$ y $1,75$, ¿cuántas iteraciones son necesarias hasta obtener, mediante el método de bisección, un intervalo de amplitud menor o igual que $0,0001$ que contenga dicha raíz?

Solución. Como $a = 1{,}55$ y $b = 1{,}75$, entonces:

$$\frac{1}{2^n} \cdot (1{,}75 - 1{,}55) \leq 0{,}0001 \Leftrightarrow 2^n \geq \frac{0{,}2}{0{,}0001} = 2000$$

Y $n = 11$ es el primer número natural que lo verifica. Por tanto, serán necesarias 11 iteraciones para conseguir un intervalo de amplitud menor o igual que 0,0001 que contenga a la raíz.

4. Es aplicable el teorema del punto fijo a la función $g(x) = 3 - x^2$ en el intervalo $[-\frac{1}{4}, \frac{1}{4}]$.

Solución. Dicho teorema no es aplicable a la función $g(x) = 3 - x^2$ en dicho intervalo, pues para $0 \in [-\frac{1}{4}, \frac{1}{4}]$ es $g(0) = 3$ que no pertenece al mismo, por tanto no se verifica la primera de las condiciones del teorema del punto fijo.

5. Aplicando el método de Newton, partiendo de $x_0 = 0$, encontrar una raíz próxima de la ecuación

$$f(x) = 3x + \sin x - e^x = 0$$

Redondear los cálculos a cinco cifras significativas e iterar hasta que se cumpla que $|x_i - x_{i-1}| \leq 0,001$.

Solución. En este caso, nos piden simplemente aplicar el algoritmo de Newton a esta ecuación partiendo del punto $x_0 = 0$, como $f(x) = 3x + \sin x - e^x$ y $f'(x) = 3 + \cos x - e^x$.

La fórmula de iteración será:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{3x_n + \sin x_n - e^{x_n}}{3 + \cos x_n - e^{x_n}}$$

Y comenzando el proceso iterativo:

$$x_0 = 0$$
$$x_1 = 0 - \frac{3 \cdot 0 + \sin 0 - e^0}{3 + \cos 0 - e^0} = \frac{1}{3} = 0{,}33333$$
$$x_2 = 0{,}33333 - \frac{3 \cdot 0{,}33333 + \sin 0{,}33333 - e^{0{,}33333}}{3 + \cos 0{,}33333 - e^{0{,}33333}} = 0{,}36017;$$
$$|x_2 - x_1| = 0{,}02684 > 0{,}001$$
$$x_3 = 0{,}36017 - \frac{3 \cdot 0{,}36017 + \sin 0{,}36017 - e^{0{,}36017}}{3 + \cos 0{,}36017 - e^{0{,}36017}} = 0{,}36043;$$
$$|x_3 - x_2| = 0{,}00026 < 0{,}001$$

Por tanto la raíz buscada es:

$$\overline{x} = x_3 = 0{,}36043$$

6. La ecuación $f(x) = e^x - 3x^2 = 0$ posee tres raíces reales distintas r_1, r_2 y r_3, pertenecientes a los intervalos $(-1, 0)$, $(0, 1)$ y $(3, 4)$; se desea aproximar la menor de ellas por el método de Newton, la mayor por el método de Lagrange y la intermedia escribiendo la ecuación en la forma $x = g(x)$ con g adecuadamente elegida, realizando cuatro pasos en cada caso, estimando el error correspondiente a la cuarta iteración y justificando los procedimientos utilizados.

   Solución. Vamos a aplicar el **método de Newton en el intervalo** $[-1, 0]$, en primer lugar se verifican las propiedades del teorema de convergencia global-1: 1) $f(-1)f(0) < 0$, y es fácil probar 2) $f'(x) > 0$ y 3) $f''(x) < 0$ en dicho intervalo, luego existe una única raíz de $f(x) = 0$ en el mismo y si tomamos el punto inicial $x_0 = -1$, como además se verifica la 4) $f(-1)f''(-1) > 0$, el método de Newton converge a dicha raíz. Veamos las cuatro primeras iteraciones del algoritmo del método de Newton dado ahora por

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{e^{x_n} - 3x_n^2}{e^{x_n} - 6x_n}$$

   que resultan ser

$$x_1 = -0{,}58665665970203$$
$$x_2 = -0{,}46980190772452$$
$$x_3 = -0{,}45905391695502$$
$$x_4 = -0{,}45896227419484$$

Para estimar el error de esta última aproximación utilizaremos la fórmula de la proposición 2, tomando $0 < m \leq \text{mín}_{x\in[-1,0]} \mid f'(x) \mid= 1$ resultará que $\mid x_4 - r_1 \mid\leq\mid f(x_4) \mid= 2{,}2541711275358978 \cdot 10^{-8}$, luego esta aproximación tiene siete cifras correctas, lo que muestra la bondad de este método. En el caso que nos ocupa, dado que también se verifica que $\text{máx}\{\mid \frac{f(-1)}{f'(-1)} \mid, \mid \frac{f(0)}{f'(0)} \mid\} = 1 \leq b-a = 0-(-1) = 1$, el método converge, según el teorema de convergencia global-2, para cualquier punto inicial $x_0 \in [-1, 0]$.

Ahora, como nos piden, obtendremos la aproximación a la raíz en $[3, 4]$ utilizando **el método de Lagrange**. En primer lugar, es fácil ver que $f(3) = -6{,}914463076812332 < 0$ y $f(4) = 6{,}598150033144236 > 0$, $f'(x) > 0$ y $f''(x) > 0$ para todo $x \in [3, 4]$. Luego aquí el extremo fijo es el $b = 4$ y el algoritmo resulta ser

$$x_0 = 3,\ x_{n+1} = g(x_n) = \frac{x_n f(4) - 4f(x_n)}{f(4) - f(x_n)}$$

siendo ahora las cuatro primeras iteraciones de este algoritmo:

$$\begin{aligned} x_1 &= 3{,}51170436247579 \\ x_2 &= 3{,}680658256169178 \\ x_3 &= 3{,}721559745743162 \\ x_4 &= 3{,}730592116693345 \end{aligned}$$

Para estimar el error de esta cuarta iteración podemos utilizar el mismo método anterior, siendo ahora $0 < m \leq \text{mín}_{x\in[3,4]} \mid f'(x) \mid= 2{,}085536923187668$, para simplificar podemos tomar $m = 2$, resultando $\mid x_4 - r_3 \mid\leq \frac{|f(x_4)|}{2} = 0{,}024079127915652$, aquí las aproximaciones a la raíz buscada progresan más lentamente.

Finalmente, para aproximar $r_2 \in [0, 1]$, escribiendo la ecuación **en la forma equivalente** $x = g(x)$, en nuestro caso como $x = \sqrt{\frac{e^x}{3}}$, veamos en primer lugar que la función $g(x) = \sqrt{\frac{e^x}{3}} = \frac{e^{\frac{x}{2}}}{\sqrt{3}}$ verifica las dos condiciones del teorema del punto fijo, en efecto $g'(x) = \frac{e^{x/2}}{2\sqrt{3}} > 0$ para todo $x \in [0, 1]$ luego $g(x)$ es estrictamente creciente y como $g(0) = 0{,}57735026918963$ y $g(1) = 0{,}95188966945738$ se sigue que $0 < g(0) < g(x) < g(1) < 1$, por tanto se cumple la primera condición; por otro lado, como la derivada segunda de g es positiva la g' toma su máximo valor $g'(1) = 0{,}47594483472869$, luego $\mid g'(x) \mid\leq \frac{1}{2}$, que implica la segunda condición, por tanto el método iterativo definido por esta g es convergente partiendo de cualquier $x_0 \in [0, 1]$, por

ejemplo si partimos de $x_0 = 1$ resultan las aproximaciones

$$x_1 = 0{,}95188966945738$$
$$x_2 = 0{,}92926501698664$$
$$x_3 = 0{,}91881210281326$$
$$x_4 = 0{,}9140224980218$$

El error de la última aproximación obtenida puede testarse, ahora, por la fórmula $\mid x_4 - r_2 \mid \leq \frac{1/2}{1-1/2} \mid x_4 - x_3 \mid = \mid x_4 - x_3 \mid = 0{,}00478960479146$, luego en este caso podemos asegurar que esta aproximación tiene dos cifras correctas.

7. Hallar la raíz cuadrada de 10 usando tres iteraciones mediante el método de Newton y comenzando con el valor inicial $x_0 = 3$, justificando que puede utilizarse dicho método. Utilícense dos decimales redondeados en los cálculos.

Solución. Una ecuación que nos permite encontrar la raíz cuadrada de diez es:

$$x^2 - 10 = 0$$

Así, $f(x) = x^2 - 10$, $f'(x) = 2x$ y $f''(x) = 2$, luego en virtud del teorema 6 (convergencia global-2) puede utilizarse dicho método en el intervalo $[3, 4]$, comenzando a iterar por cualquier punto del mismo. La fórmula de iteración será:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 10}{2x_n}$$

y comenzando por el punto $x_0 = 3$, trabajando con dos decimales redondeados, obtenemos las aproximaciones:

$$x_1 = 3 - \frac{3^2 - 10}{2 \cdot 3} = 3,17$$
$$x_2 = 3,17 - \frac{3,17^2 - 10}{2 \cdot 3,17} = 3,16$$
$$x_3 = 3,16 - \frac{3,16^2 - 10}{2 \cdot 3,16} = 3,16$$

Por tanto

$$\sqrt{10} \simeq 3,16$$

con dos cifras decimales redondeadas. Si hubiésemos trabajado con más cifras decimales obtendríamos con el mismo número de pasos esta raíz

con una aproximación bastante mejor, por ejemplo trabajando con Maxima con 15 cifras decimales, obtendríamos la raíz con once cifras decimales correctas en sólo tres pasos.

8. Asimismo, probar que la ecuación $x = cosx$ posee una única raíz que pertenece al intervalo $[0, 1]$ y que puede aproximarse por el método de Newton, obtener una aproximación de la misma con error menor que $5 \cdot 10^{-6}$ y comprobar la convergencia cuadrática del método en este caso.

   Solución. Sea $f(x) = x - \cos x$, puesto que $f'(x) = 1 + \sin x \geq 1 > 0$ para todo $x \in \mathbb{R}$ la función es estrictamente creciente, luego posee a lo más una raíz, como $f(0) = -1$ y $f(1) = 1-\cos 1 > 0$ se sigue que posee una única raíz r en el intervalo $[0, 1]$. Por otro lado, $f''(x) = \cos x > 0$ para todo $x \in [0, 1]$, por tanto el método de Newton converge partiendo del punto $x_0 = 1$ pues $f(x_0) \cdot f'(x_0) > 0$ (también se verifica el teorema de convergencia global-2 y el método converge para cualquier valor inicial $x_0 \in [0, 1]$)a dicha raíz. Así pues, partiendo de $x_0 = 1$ y aplicando el algoritmo de Newton:

$$x_{n+1} = x_n - \frac{x_n - \cos x_n}{1 + \sin x_n}$$

   se obtienen las aproximaciones:

$$\begin{aligned} x_1 &= 0{,}750363867840241 \\ x_2 &= 0{,}73911289091136 \\ x_3 &= 0{,}73908513338528 \\ x_4 &= 0{,}73908513321516 \end{aligned}$$

   Ahora bien se verifica que $0 < m = 1 = \text{mín}_{x\in[0,1]} \mid f'(x) \mid$, y el error absoluto es $\mid x_4 - r \mid \leq \mid f(x_4) \mid = 1{,}1102230246251565 \cdot 10^{-15} < 5 \cdot 10^{-6}$, puede comprobarse por este mismo procedimiento que también x_3 cumple el requisito ya que $\mid x_3 - r \mid \leq \mid f(x_3) \mid = 2{,}8471391910755983 \cdot 10^{-10} < 5 \cdot 10^{-6}$.

   La convergencia es cuadrática en este caso pues $f''(x) \neq 0$ para todo $x \in [0, 1]$, por tanto en la raíz r será $g''(r) = \frac{f''(r)}{f'(r)} \neq 0$.

9. Dada la ecuación: $e^x - 1{,}5 - \arctan x = 0$, probar que posee una única raíz positiva, expresarla en la forma $x = g(x)$, con g adecuadamente elegida por vosotros en algún intervalo conteniendo a dicha raíz para que, partiendo de $x_0 = 1$, el método $x_{n+1} = g(x_n)$ sea convergente, aproximarla con error menor que $5 \cdot 10^{-5}$. Hallarla, con la misma cota

de error, por el método de Newton (justificando su aplicación). ¿Cuál es el orden exacto de ambos métodos en este caso?.

Solución. Sea $f(x) = e^x - 1{,}5 - \arctan x$, entonces $f'(x) = e^x - \frac{1}{1+x^2}$ y $f''(x) = e^x + \frac{2x}{(1+x^2)^2}$, puesto que $f''(x) > 0$ para toda $x \geq 0$ y $f'(0) = 0$ se sigue que $f'(x) > 0$ para toda $x > 0$, luego f es estrictamente creciente en $(0, \infty)$, luego posee a lo más una raíz positiva; por otro lado, $f(0) = 1 - 1{,}5 = -0{,}5 < 0$ y $f(1) = e - 1{,}5 - arctan 1 = 0{,}4328836650616 > 0$, por tanto existe una única raíz positiva que pertenece al intervalo $[0, 1]$, como también es $f(0{,}5) = -0{,}31492633830068 < 0$, dicha raiz pertenece al intervalo [0.5,1].

Vamos a aproximar dicha raíz de dos formas:

1ª) Para la primera, escribimos la ecuación $f(x) = 0$ en la forma equivalente $e^x = 1{,}5 + \arctan x$ o también $x = \log(1{,}5 + \arctan x) = g(x)$ (aquí log denota el logaritmo neperiano). Veamos que $g(x)$ verifica las dos condiciones del teorema del punto fijo en el intervalo [0.5,1]. En efecto, $g(0{,}5) = 0{,}67480376868132$, $g(1) = 0{,}82654026010604$, puesto que $g'(x) = \frac{1}{(1+x^2)(1{,}5+\arctan x)} > 0$ para todo x en [0.5,1], la g es estrictamente creciente en dicho intervalo y por tanto se verifica que para todo x del intervalo [0.5,1] es $0{,}5 < g(0{,}5) < g(x) < g(1) < 1$ o sea $g(x)$ pertenece a [0.5,1], que es la primera condición del teorema del punto fijo. Ahora, es fácil probar que para todo x de [0.5,1] es $\mid g'(x) \mid \leq \frac{1}{(1+0{,}5)(1{,}5+\arctan(0{,}5))} \leq 0{,}4074050742776 < \frac{1}{2} < 1$, lo cual implica la segunda condición requerida en dicho teorema. Así pues, g posee un único punto fijo r en el intervalo [0.5,1] y se puede aplicar el método iterativo definido por g, partiendo de cualquier punto inicial en dicho intervalo, si tomamos el punto inicial $x_0 = 1$ como nos piden, se obtienen las aproximaciones:

$$
\begin{aligned}
x_1 &= 0{,}82654026010604 \\
x_2 &= 0{,}78422835461746 \\
x_3 &= 0{,}77244337638646 \\
x_4 &= 0{,}76904785609329 \\
x_5 &= 0{,}76806015379644 \\
x_6 &= 0{,}76777205408692 \\
x_7 &= 0{,}76768795174874 \\
x_8 &= 0{,}76766339476652
\end{aligned}
$$

Podemos comprobar que esta última aproximación x_8 verifica el requisito de ser $\mid x_8 - r \mid \leq \frac{1/2}{1-1/2} \mid x_8 - x_7 \mid = \mid x_8 - x_7 \mid =$

$2{,}4556982220036438 \cdot 10^{-5} < 5 \cdot 10^{-5}$. Como $0{,}21 <\mid g'(x) \mid < 0{,}41 < 1$ el método utilizado es de convergencia lineal.

2ª) Utilizamos ahora el método de Newton, se ha puesto de manifiesto al comienzo del ejercicio que se cumplen todas las condiciones del teorema de convergencia global-1 en el intervalo [0.5,1], si comenzamos a iterar por $x_0 = 1$, con el algoritmo:

$$x_{n+1} = x_n - \frac{e^{x_n} - 1{,}5 - \arctan x_n}{e^{x_n} - \frac{1}{1+x_n{}^2}}$$

obtendremos las aproximaciones

$$\begin{aligned} x_1 &= 0{,}804856326410831 \\ x_2 &= 0{,}7688453128309486 \\ x_3 &= 0{,}7676545507181154 \end{aligned}$$

Tomando ahora $m = \min_{x\in[0{,}5{,}1]} \mid f'(x) \mid = 0{,}84872127070013$, el error absoluto de x_3 será $\mid x_3 - r \mid \leq \frac{|f(x_3)|}{0{,}84872127070013} < 2{,}5 \cdot 10^{-6}$, que verifica la condición requerida. La convergencia es exactamente cuadrática, en este caso, pues $f''(x) \neq 0$ para todo x en [0.5,1], por tanto $g''(r) = \frac{f''(r)}{f'(r)} \neq 0$.

10. Dada la ecuación $(1+x)^2 \sin x - x - 1 = 0$, probar que posee una única solución en el intervalo $[\frac{\pi}{8}, \frac{\pi}{2}]$, expresarla en la forma $x = g(x)$ con g adecuadamente elegida en dicho intervalo para poder aplicar el teorema del punto fijo; partiendo del punto medio de dicho intervalo, obtener x_6 y estimar el error de dicha aproximación.

Solución. La ecuación propuesta es equivalente en $[\frac{\pi}{8}, \frac{\pi}{2}]$ a las siguientes

$$(1+x)^2 \sin x = 1 + x \Leftrightarrow \sin x = \frac{1}{1+x} \Leftrightarrow x = \arcsin \frac{1}{1+x}$$

ahora, si definimos $g(x) = \arcsin \frac{1}{1+x}$, veamos que se cumplen las dos condiciones del teorema del punto fijo, en primer lugar

$$g'(x) = \frac{-1}{(x+1)\sqrt{x^2+2x}}$$

para todo x del intervalo es obvio que $g'(x) < 0$, por tanto g es estrictamente decreciente en dicho intervalo y además

$$\mid g'(x) \mid \leq \frac{1}{(1+\frac{\pi}{8})\sqrt{(\frac{\pi}{8})^2 + \frac{\pi}{4}}} < 0{,}75 = k$$

pues para $x = \frac{\pi}{8}$ el denominador es menor y el cociente mayor, por otro lado $g(\frac{\pi}{8}) = 0{,}800968...$, $g(\frac{\pi}{2}) = 0{,}399529...$, y al ser g estrictamente decreciente en $[\frac{\pi}{8}, \frac{\pi}{2}]$ se tiene

$$\frac{\pi}{2} > g(\frac{\pi}{8}) > g(x) > g(\frac{\pi}{2}) > \frac{\pi}{8}$$

luego para todo $x \in [\frac{\pi}{8}, \frac{\pi}{2}] \Rightarrow g(x) \in [\frac{\pi}{8}, \frac{\pi}{2}]$, que junto con la anterior son las dos condiciones que garantizan que posee una única raíz, r, y que el método converge para todo x_0 en dicho intervalo. Partiendo de $x_0 = (\pi/8 + \pi/2)/2$ obtenemos las seis primeras aproximaciones:

$$\begin{aligned} x_1 &= g(x_0) = 0{,}52892450088578 \\ x_2 &= g(x_1) = 0{,}71293206563925 \\ x_3 &= g(x_2) = 0{,}62339424844312 \\ x_4 &= g(x_3) = 0{,}66364629202218 \\ x_5 &= g(x_4) = 0{,}64486344137671 \\ x_6 &= g(x_5) = 0{,}65348012238029 \end{aligned}$$

El error absoluto de la última aproximación puede acotarse en la forma

$$\mid x_6 - r \mid \leq \frac{0{,}75}{1 - 0{,}75} \mid 0{,}65348012238029 - 0{,}64486344137671 \mid \leq 0{,}026$$

luego podemos asegurar que, al menos, es correcta la primera cifra decimal de x_6.

11. Aplicar el método de la secante para encontrar una raíz de la ecuación $\cos x - x = 0$ en el intervalo $[0{,}5, \pi/4]$. Trabajando con tres decimales redondeados e iterando hasta que se verifique que $|x_i - x_{i-1}| \leq 0,005$.

Solución. Sean

$$f(x) = \cos x - x, \; x_0 = 0{,}5, \; x_1 = \pi/4 = 0{,}785$$

Entonces $f(0,5) = 0,378$ y $f(0,785) = -0,078$.

La fórmula iterativa del método de la secante es:

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)}$$

Por tanto,

$$x_2 = x_1 - f(x_1) \cdot \frac{x_0 - x_1}{f(x_0) - f(x_1)} = 0,785 + 0,078 \cdot \frac{0,5 - 0,785}{0,378 + 0,078} = 0,736$$

Como $f(0,736) = 0,005$, entonces

$$x_3 = x_2 - f(x_2) \cdot \frac{x_1 - x_2}{f(x_1) - f(x_2)} = 0,736 - 0,005 \cdot \frac{0,785 - 0,736}{-0,078 - 0,005} = 0,739$$

$$|x_3 - x_2| = |0,739 - 0,736| = 0,003 < 0,005$$

Así la aproximación de la raíz en dicho intervalo es

$$\overline{x} = x_3 = 0,739$$

12. Hallar una raíz aproximada en el intervalo $[1,2]$ de la ecuación $x^3 - x + 1 = 0$, utilizando en los cálculos cuatro decimales redondeados e iterar hasta que $|x_i - x_{i-1}| \leq 0,1 \cdot 10^{-2}$. Primero, por el método de Newton y luego por el método de la secante.

Solución. Aplicando el método de Newton:

Como $f(x) = x^3 - x - 1$ y $f'(x) = 3x^2 - 1$.

La fórmula de iteración utilizada será:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^3 - x_n - 1}{3x_n^2 - 1}$$

Y comenzando el proceso iterativo:

$$x_0 = 1$$

$$x_1 = 1 - \frac{1^3 - 1 - 1}{3 \cdot 1^2 - 1} = 1{,}5$$

$$x_2 = 1{,}5 - \frac{1{,}5^3 - 1{,}5 - 1}{3 \cdot 1{,}5^2 - 1} = 1{,}3478; |x_2 - x_1| = 0{,}1522 > 0{,}1 \cdot 10^{-2}$$

$$x_3 = 1{,}3478 - \frac{1{,}3478^3 - 1{,}3478 - 1}{3 \cdot 1{,}3478^2 - 1} = 1{,}3252; |x_3 - x_2| = 0{,}0226$$

$$x_4 = 1{,}3252 - \frac{1{,}3252^3 - 1{,}3252 - 1}{3 \cdot 1{,}3252^2 - 1} = 1{,}3247; |x_3 - x_2| = 0{,}0005$$

Por tanto la solución aproximada es:

$$\overline{x} = x_4 = 1{,}3247$$

Aplicando el método de la secante:

La fórmula iterativa del método de la secante es:

$$x_{n+1} = x_n - f(x_n) \cdot \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)}$$

donde $f(x)$ es la misma que en el apartado anterior.

Tomando como valores iniciales los extremos del intervalo, es decir, $x_0 = 1$ y $x_1 = 2$. Entonces $f(1) = -1$ y $f(2) = 5$. Así:

$$x_2 = x_1 - f(x_1) \cdot \frac{x_0 - x_1}{f(x_0) - f(x_1)} = 2 - 5 \cdot \frac{1-2}{-1-5} = 1{,}1667;\ f(1{,}667) = -0{,}5786$$

Como $f(0,736) = 0,005$, entonces

$$x_3 = x_2 - f(x_2) \cdot \frac{x_1 - x_2}{f(x_1) - f(x_2)} = 1{,}1667 + 0{,}5786 \cdot \frac{2 - 1{,}1667}{5 + 0{,}5786} = 1{,}2531;$$

y siendo ahora $f(1{,}2531) = -0{,}2854$ obtendremos

$$x_4 = 1{,}2531 + 0{,}2854 \cdot \frac{1{,}1667 - 1{,}2531}{-0{,}5786 + 0{,}2854} = 1{,}3372;\ f(1{,}3372) = 0{,}0539$$

$$x_5 = 1{,}3372 - 0{,}0539 \cdot \frac{1{,}2531 - 1{,}3372}{-0{,}2854 - 0{,}0539} = 1{,}3238;\ f(1{,}3238) = -0{,}0039;$$

$$|x_5 - x_6| = 0{,}0134 > 0{,}001$$

$$x_6 = 1{,}3238 + 0{,}0039 \cdot \frac{1{,}3372 - 1{,}3238}{0{,}0539 + 0{,}0039} = 1{,}3247;\ f(1{,}2531) = -0{,}2854;$$

$$|x_6 - x_7| = 0{,}0004 < 0{,}001$$

Así la aproximación de la raíz en dicho intervalo es

$$\overline{x} = x_6 = 1{,}3247$$

Como puede observarse, el método de la secante es más lento en su convergencia que el método de Newton. Tiene, sin embargo, la ventaja de que no utiliza derivadas, que en algunas funciones pueden ser complejas en su cálculo y aplicación.

13. Se considera la ecuación de Kepler: $f(x) = 0{,}5 - x + 0{,}2 \cdot sinx = 0$, se desea determinar un intervalo y una función de iteración tal que pueda aplicarse el teorema del punto fijo para aproximar su raíz con error menor que $5 \cdot 10^{-5}$; hacerlo también por el método de Newton, justificando en ambos casos los métodos utilizados.

Solución. Dada la ecuación $f(x) = 0{,}5 - x + 0{,}2 \sin x = 0$ es inmediato ver que $f(0) = 0{,}5 > 0$ y que $f(1) = 0{,}5 - 1 + 0{,}2 \sin 0{,}5 = -0{,}33170580303842 < 0$, luego tiene al menos una raíz en $[0, 1]$, puesto que la derivada $f'(x) = -1 + 0{,}2 \cos x < 0$ para todo x, se deduce que la ecuación posee una única raíz r en dicho intervalo.

La ecuación dada puede escribirse en la forma equivalente $x = 0{,}5 + 0{,}2 \sin x = g(x)$, donde $g'(x) = 0{,}2 \cos x > 0$ en [0,1] y por tanto g es estrictamente creciente en dicho intervalo, siendo el módulo de la derivada menor o igual a 0,2, se ve fácilmente que se cumplen las condiciones de convergencia para todo punto inicial en [0,1], pues para todo x del intervalo $[0, 1]$: 1) $0 < g(0) < g(x) < g(1) < 1$ y 2) $\mid g'(x) \mid \leq 0{,}2 = k < 1$. Comenzando a iterar por $x_0 = 1$, mediante el algoritmo $x_{n+1} = g(x_n)$ obtenemos las aproximaciones siguientes:

$$\begin{aligned} x_1 &= 0{,}66829419696158 \\ x_2 &= 0{,}62392960776892 \\ x_3 &= 0{,}61684577785978 \\ x_4 &= 0{,}61569302419112 \\ x_5 &= 0{,}61550488548918 \\ x_6 &= 0{,}61547416515862 \end{aligned}$$

Cumpliendo el error absoluto de la sexta iteración la desigualdad:

$$\mid x_6 - r \mid \leq \frac{0{,}2^6}{1 - 0{,}2} \mid x_1 - x_0 \mid = 2{,}6536464243073669 \cdot 10^{-5} < 5 \cdot 10^{-5}$$

como se requería en el enunciado.

Resolvamos nuevamente este problema por el método de Newton, vemos que en el intervalo $[0, 1]$ se cumplen las propiedades: 1) $f(0)f(1) < 0$, 2) $f'(x) < 0$, 3) $f''(x) = -0{,}2 \sin x \leq 0$ y 4) $f(1)f''(1) > 0$, luego el método de Newton converge partiendo del punto $x_0 = 1$ (de hecho converge para cualquier punto inicial en el intervalo $[0, 1]$ pues también se cumple la otra propiedad), obteniéndose las tres primeras aproximaciones que siguen:

$$\begin{aligned} x_1 &= 0{,}62810730032791 \\ x_2 &= 0{,}61547930374494 \\ x_3 &= 0{,}61546816949852 \end{aligned}$$

Puesto que el $\text{mín}_{x \in [0,1]} \mid f'(x) \mid = 0{,}8$ el error absoluto de la tercera aproximación puede acotarse en la forma $\mid x_3 - r \mid \leq \frac{|f(x_3)|}{0{,}8} \leq$

$8{,}9468363273503826 \cdot 10^{-12}$, que verifica sobradamente el requisito, que puede comprobarse también lo verifica x_2.

14. Sea r una raíz de multiplicidad p (entero mayor que 1) de $f(x) = 0$ (es decir $f(x) = (x-r)^p h(x)$, con $h(r) \neq 0$), probar que si f es de clase $\mathcal{C}^{(3)}$ en un entorno de r, el método de Newton es de orden uno, y que si se modifica en la forma

$$x_{n+1} = x_n - p \cdot \frac{f(x_n)}{f'(x_n)}$$

entonces tiene convergencia, al menos, cuadrática.

Solución. En este caso se tiene que

$$f'(x)) = p(x-r)^{p-1}h(x) + (x-r)^p h'(x)$$

y la función de iteración del método de Newton

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

se reduce a

$$g(x) = x - \frac{(x-r)^p h(x)}{p(x-r)^{p-1}h(x) + (x-r)^p h'(x)} = x - \frac{(x-r)h(x)}{ph(x) + (x-r)h'(x)}$$

siendo esta función de clase $\mathcal{C}^{(2)}$, por ser f de clase $\mathcal{C}^{(3)}$, en tanto que para su derivada obtenemos

$$\begin{aligned} g'(x) = 1 - [(h(x) + (x-r)h'(x))(ph(x) + (x-r)h'(x)) - ((p+1)h'(x) + \\ +(x-s)h''(x))(x-r)h(x)]/(ph(x) + (x-r)h'(x))^2 \end{aligned}$$

Ahora, haciendo $x = r$ se obtiene

$$g'(r) = 1 - \frac{ph(r)^2}{p^2 h(r)^2} = 1 - \frac{1}{p}$$

y como $0 < g'(r) = 1 - \frac{1}{p} < 1$, se deduce que la convergencia del método de Newton es ahora lineal o de primer orden. Pero si se modifica el método en la forma $x_{n+1} = x_n - p \cdot \frac{f(x_n)}{f'(x_n)}$, entonces se tendrá que $g'(r) = 1 - p \cdot \frac{1}{p} = 0$, lo que implica en este caso que la convergencia es, al menos, cuadrática.

(Otra forma de resolverlo es la siguiente, si $x = r$ es una raíz de orden p de $f(x) = 0$ entonces es una raíz de orden uno de la ecuación $f(x)^{\frac{1}{p}} = 0$ y aplicando el método de Newton a esta ecuación se obtiene el algoritmo del método de Newton modificado con convergencia, al menos, cuadrática para raíces múltiples de la ecuación de partida).

15. Dada la ecuación: $x^3 + 2x^2 + 10x - 20 = 0$, se pide:

    - Acotar sus raíces reales y probar que tiene une única raíz real en el intervalo $[1, 2]$.
    - Expresarla en la forma $x = g(x)$, siendo $g(x)$ una fracción tal que genere un método iterativo convergente en $[1, 2]$, hacer tres iteraciones partiendo de $x_0 = 1{,}5$.
    - Aproximar la raíz anterior por el método de Newton con error menor que $5 \cdot 10^{-6}$, justificando la convergencia del mismo y el cumplimiento de la cota del error.

    Solución. La primera parte la haremos por el método de los agrupamientos, como en $p(x) = x^3 + 2x^2 + 10x - 20 = 0$ hay un único grupo y es positivo para $\xi_0 = 2$, se sigue que $L = 2$ es una cota superior de sus raíces reales. Ahora, cambiamos x por $-x$ y resulta $p(-x) = -x^3+2x^2-10x-20$ y cambiando de signo, para que comience por signo positivo obtenemos la ecuación $-p(-x) = x^3-2x^2+10x+20$, cuyas raíces son opuestas a la de partida, esta tiene dos grupos: x^3-2x^2, que es positivo para $\xi_0 = 3$ y $10x+20$ que lo es, por ejemplo, para $\xi_1 = 1$, por tanto $L' = 3$ es una cota superior de sus raíces reales y por consiguiente siendo estas opuestas a las de la ecuación de partida, si r es una raíz real de la ecuación de partida es $-r < L'$ por tanto $l = -L' < r$, es decir toda raíz real de la ecuación dada verifica que $l = -L' < r < L$ o sea $-3 < r < 2$, con lo que hemos acotado las raíces reales de la ecuación dada. Veamos que tiene una única raíz real en $[1, 2]$, en efecto $p(1) = -7 < 0$, $p(2) = 16 > 0$ y por continuidad de $p(x)$ posee, al menos, una raíz real en $[1, 2]$; ahora $p'(x) = 3x^2 + 4x + 10 > 0$ para todo $x \in [1, 2]$ ya que es suma de números positivos por tanto $p(x)$ es estrictamente creciente en dicho intervalo y por consiguiente posee una única raíz real en el mismo.

    Para la segunda parte, vemos que la ecuación dada es equivalente en $[1, 2]$ a la siguiente

$$x^3+2x^2+10x = 20 \Leftrightarrow x(x^2+2x+10) = 20 \Leftrightarrow x = \frac{20}{x^2 + 2x + 10} = g(x)$$

veamos a continuación que esta g verifica las dos condiciones del teorema de la aplicación contractiva en $[1, 2]$:

i) $\forall x \in [1, 2]$ el denominador $x^2 + 2x + 10$ es creciente y varia entre 13 y 18 por tanto

$$1 < 20/18 = 1{,}111\ldots < \frac{20}{x^2 + 2x + 10} < 20/13 = 1{,}538\ldots < 2$$

es decir $g(x) \in [1, 2]$.

ii) También puede probarse que $g'(x) = \frac{-40(x+1)}{(x^2+2x+10)^2}$ y su módulo resulta ser

$$\mid g'(x) \mid = \frac{40(x+1)}{(x^2 + 2x + 10)^2} \leq \mid g'(1) \mid = 0{,}47\ldots < 0{,}5$$

lo que implica la segunda condición.

Por tanto, el método iterativo definido por esta función g converge a la única raíz r en $[1, 2]$ partiendo de cualquier punto x_0 en dicho intervalo, tomando $x_0 = 1{,}5$, obtenemos las tres primeras aproximaciones de r que siguen:

$$x_1 = 1{,}311475409836066$$
$$x_2 = 1{,}394416338767098$$
$$x_3 = 1{,}357475620650471$$

Si se desea se puede estimar el error absoluto de la última aproximación hallada mediante la fórmula

$$\mid x_3 - r \mid \leq \frac{0{,}5}{1 - 0{,}5} \mid x_3 - x_2 \mid = \mid x_3 - x_2 \mid = 0{,}036940718116627$$

que nos asegura tan sólo que es correcta la primera cifra decimal de la tercera aproximación.

Finalmente, apliquemos el método de Newton $x_{n+1} = x_n - \frac{p(x_n)}{p'(x_n)}$, puesto que 1) $p(1)p(2) < 0$, 2)$p'(x) > 0$ para todo $x \in [1, 2]$, 3) $p''(x) > 0$ para todo $x \in [1, 2]$ y 4) $p(1{,}5)p'(1{,}5) > 0$; el método converge a la única raíz r, obteniendo las tres primeras aproximaciones que siguen

$$x_1 = 1{,}373626373626374$$
$$x_2 = 1{,}368814819623964$$
$$x_3 = 1{,}368808107834412$$

Veamos si x_3 verifica el requisito de error absoluto, para ello basta con tomar m tal que $0 < m \leq \text{mín}_{x\in[1,2]} \mid p'(x) \mid= 17$, con lo que podemos acotar el error absoluto en la forma

$$\mid x_3 - r \mid \leq \frac{\mid p(x_3) \mid}{17} = 1{,}6181356907990824 \cdot 10^{-11} < 5 \cdot 10^{-6}$$

que muestra no sólo que esta aproximación cumple el requisito pedido, sino que las diez primeras cifras decimales de esta aproximación de la raíz son correctas.

16. Efectuar dos iteraciones por el método de Newton, partiendo del punto $(0,1)$, para aproximar la solución del sistema:

$$4x^2 - y^2 = 0, \; 4xy^2 - x - 1 = 0$$

Solución. Como hemos visto en la parte teórica el método de Newton para sistemas no lineales consiste en partir de una aproximación inicial dada $\bar{x}^{(0)}$, y realizar el algoritmo siguiente:

$$\begin{array}{c}\text{Hallar } \delta^{(m)} \text{ verificando } J_f(\bar{x}^{(m)})\delta^{(m)} = -f(\bar{x}^{(m)}) \\ \text{y obtener } \bar{x}^{(m+1)} = \bar{x}^{(m)} + \delta^{(m)}\end{array}$$

hasta que $\delta^{(m)}$ sea suficientemente pequeño.

En el caso que nos ocupa, nos piden dos iteraciones partiendo de $\bar{x}^{(0)} = (0,1)$, siendo $f(\bar{x}) = f(x,y) = (4x^2-y^2, 4xy^2-x-1)$, en primer lugar hemos de hallar el jacobiano de f en un punto cualquiera $\bar{x} = (x,y)$, dado por la matriz

$$J_f(x,y) = \begin{pmatrix} \frac{\partial f_1(x,y)}{\partial x} & \frac{\partial f_1(x,y)}{\partial y} \\ \frac{\partial f_2(x,y)}{\partial x} & \frac{\partial f_2(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 8x & -2y \\ 4y^2-1 & 8xy \end{pmatrix}$$

Para la primera iteración, resolvemos el sistema

$$J_f(0,1) \cdot \delta^{(0)} = -f(0,1)$$

que resulta ser

$$\begin{pmatrix} 0 & -2 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} \delta_x^{(0)} \\ \delta_y^{(0)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

cuya solución es $\delta^{(0)} = (\frac{1}{3}, \frac{-1}{2})$, por tanto tendremos

$$\bar{x}^{(1)} = \bar{x}^{(0)} + \delta^{(0)} = (0,1) + (\frac{1}{3}, \frac{-1}{2}) = (\frac{1}{3}, \frac{1}{2})$$

Para la segunda iteración, resolvemos el sistema

$$J_f(\frac{1}{3},\frac{1}{2})\cdot\delta^{(1)}=-f(\frac{1}{3},\frac{1}{2})$$

que resulta ser

$$\left(\begin{array}{cc}\frac{8}{3} & -1\\ 0 & \frac{4}{3}\end{array}\right)\left(\begin{array}{c}\delta_x^{(1)}\\ \delta_y^{(1)}\end{array}\right)=\left(\begin{array}{c}\frac{-7}{36}\\ 1\end{array}\right)$$

y cuya solución es $\delta^{(1)}=(\frac{5}{24},\frac{1}{4})$, por tanto tendremos

$$\bar{x}^{(2)}=\bar{x}^{(1)}+\delta^{(1)}=(\frac{1}{3},\frac{1}{2})+(\frac{5}{24},\frac{1}{4})=(\frac{13}{24},\frac{5}{4})=(0{,}54166666666667,1{,}25)$$

que es la segunda aproximación buscada.

2.6. Algunos programas Maxima para resolver ecuaciones no lineales

2.6.1. Método de bisección

Veamos un ejemplo de programación con block, creado a partir de otro de J. Rafael Rodríguez Galván, que realiza una utilización más elaborada del comando block para obtener los ceros aproximados de una función continua que cambia de signo en un intervalo. Aparecen en él otros comandos como condicionales y similares propios de programación más avanzada, cuyo significado el lector que conozca el método de bisección será capaz de comprender sin dificultad

```
(%i1)  biseccion(f,a,b,epsilon):=
       block(
       [numer],numer:true,
       if sign(f(a)) = sign(f(b)) then
       (print("ERROR, f tiene el mismo signo en a
       y en b"), return(false) )
       else do ( c:(a+b)/2, if( abs(f(c))<=
       2.2204460492503131*10^-16)then return(print
       ("La raiz exacta para wxmaxima es c = ",c)),
       if (b-a<epsilon)then return(print("La raiz
       aproximada es ", (a+b)/2," y f(",(a+b)/2,") = ",
       f((a+b)/2)))
       else if (sign(f(a)) = sign(f(c)))
       then a:c else b:c
       )
       )$
```

Este programa aproxima la raíz de una ecuación $f(x) = 0$, cuando f es continua y toma signos opuestos en los extremos del intervalo $[a, b]$, si f toma el mismo signo sale con un mensaje de error; deteniendo la ejecución cuando el valor absoluto de la función en un punto medio de algún intervalo generado en el proceso de bisección sea menor que el épsilon de máquina ($2{,}2204460492503131 \cdot 10^{-16}$)(en cuyo caso llamaremos a este punto la raíz exacta para wxmaxima) o cuando la longitud del enésimo subintervalo generado en el proceso sea menor que el épsilon que hayamos escogido (salida que llamaremos la raíz aproximada). Seguidamente lo aplicamos para aproximar la raíz real de la ecuación $x^3 - x - 1 = 0$ en el intervalo $[1, 2]$ con error menor que $5 \cdot 10^{-10}$.

```
(%i2)  kill(f)$ f(x):= x^3-x-1$ biseccion(f,1,2,5*10^(-10))$
```

La raíz aproximada es 1,324717957293615 y $f(1{,}324717957293615) = 2{,}0840706937974574 \cdot 10^{-10}$.

Ahora, vamos a calcular una solución aproximada de $x^6 + x - 5 = 0$ en el intervalo $[0, 2]$ con error menor que $E = 10^{-6}$, aplicando el método de bisección pero determinando el número máximo de pasos a realizar, lo presentamos por medio del siguiente block

```
(%i5) biseccion1(f,a,b,E):= block(
      [numer],numer:true,
      log2(x):=log(x)/log(2),
      nmaxpasos:entier(log2((b-a)/E))+1,
      for i:1 thru nmaxpasos do
      (
      c:(a+b)/2,
      if abs(f(c))<2*10^(-16) then
      (print("La solucion exacta de máquina es ",return(c)))
      else( if f(a)*f(c)<0 then b:c else a:c)
      ),
      print("La aproximación buscada es c = ",c)
      )$
```

```
(%i6) kill(f)$ f(x):= x^6+x-5$ biseccion1(f,0,2,10^-6)$
```

La aproximación buscada es $c = 1{,}246628761291504$

2.6.2. Método de Newton-Raphson

Sea la ecuación $e^{-x} - x = 0$, queremos aproximar la solución de dicha ecuación en el intervalo $[0, 1]$. En primer lugar, veamos que podemos aplicar el teorema de convergencia global-2 del método de Newton y por tanto que podemos comenzar a iterar por cualquier punto del intervalo dado, por ejemplo por el punto $x0 = 0{,}5$. Para ello, comencemos escribiendo la función y calculando sus derivadas primera y segunda

```
(%i9) f(x):=%e^(-x)-x;define(Df(x),diff(f(x),x));
      define(D2f(x),diff(f(x),x,2));
      Df(0);Df(1.0);
```

(%o9) $\mathrm{f}(x) := e^{-x} - x$

(%o10) $\mathrm{Df}(x) := -e^{-x} - 1$

(%o11) $\mathrm{D2f}(x) := e^{-x}$

(%o12) -2

(%o13) $-1{,}367879441171442$

```
(%i14) max(abs(f(0)/Df(0)),abs(f(1)/Df(1)));
```

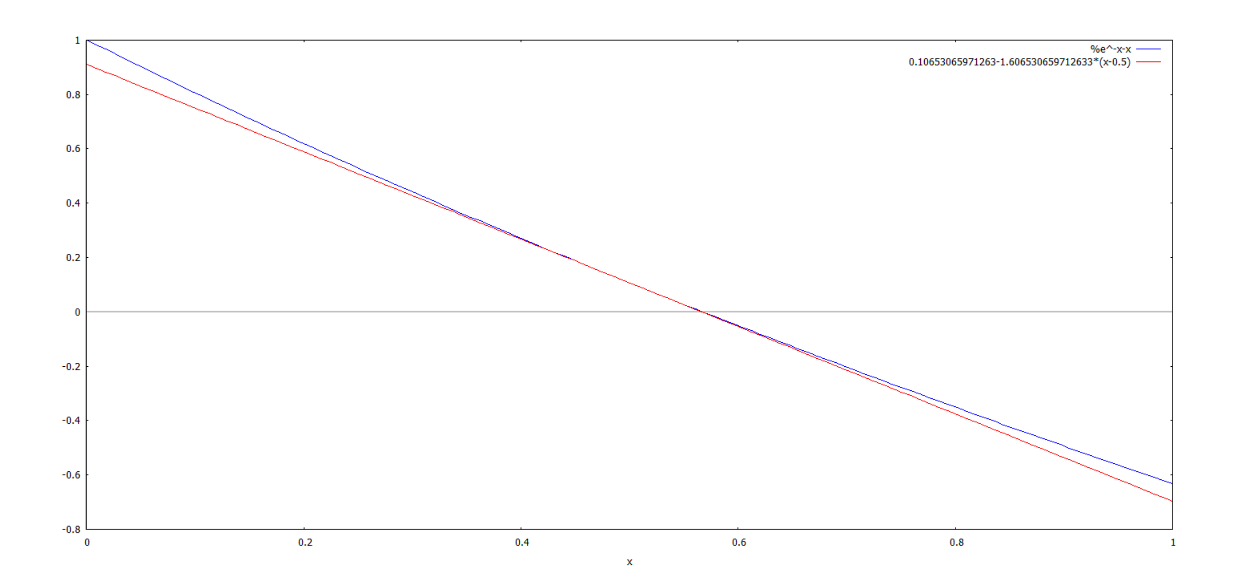
(%o14) $\frac{1}{2}$

Puesto que la derivada segunda es positiva en todo el intervalo la primera

es estrictamente creciente, ahora bien como $Df(1) < 0$ se sigue que la derivada primera es menor que 0 en todo el intervalo $[0, 1]$, por tanto $f(x) = 0$ posee una única raíz en $[0, 1]$; además, como $max|f(0)/Df(0)|, |f(1)/Df(1)| = 1/2 < 1$, el método de Newton converge partiendo de cualquier valor inicial en el intervalo, tomemos pues $x0 = 0{,}5$. Calculemos la recta tangente a la función en $x0$ y dibujemos la función y la recta en el intervalo $[0, 1]$

```
(%i15) x0:0.5$ r0(x):=f(x0)+Df(x0)*(x-x0);
       plot2d([f(x),r0(x)],[x,0,1]);
```

(%o16) $\mathrm{r0}(x) := \mathrm{f}(x0) + \mathrm{Df}(x0)\,(x - x0)$
(%o17)



Y el punto de corte de la recta tangente $r0(x)$ con el eje OX resulta ser

```
(%i18) x1:x0-f(x0)/Df(x0);
```

(%o18) 0,56631100319722

Para hallar la segunda iteración se hace lo mismo partiendo del punto x1, se obtiene así x2, luego partiendo de este del mismo modo obtenemos x3, que resultan ser

```
(%i19) x2:x1-f(x1)/Df(x1);
       x3:x2-f(x2)/Df(x2);
```

(%o19) 0,56714316503486

(%o20) 0,56714329040978

Finalmente, podemos acotar el error absoluto de $x3$ en la forma $|x3 - r| <= |f(x3)|/m$, siendo r la raíz exacta buscada y $0 < m \leq |f'(x)|$ para todo x en $[0, 1]$, puede tomarse $m = 1{,}3$.

```
(%i21) print("El error absoluto de x3 <= ",abs(f(x3)/1.3))$
```

El error absoluto de $|x3 - r| \leq 3{,}4160708450004814 \cdot 10^{-15}$.

Supongamos que se cumplen condiciones suficientes para aplicar el método de Newton para aproximar la raíz de una ecuación $f(x) = 0$ en un intervalo $[a, b]$, partiendo de un punto $x0$ (lo cual se puede estudiar a priori, de acuerdo con los resultados vistos), vamos a escribir una función que nos de una aproximación de la raíz parando el programa cuando se sobrepase el número máximo de iteraciones previstas o bien cuando la diferencia entre dos aproximaciones consecutivas sea menor que un cierto épsilon dado.

```
(%i22) newton(f,x0,epsilon,nmax):= block(
       [numer],numer:true,x[0]:x0,
       define(Df(x),diff(f(x),x)),
       for i:1 thru nmax do
       (
       x[i]:x[i-1]-f(x[i-1])/Df(x[i-1]),
       if abs(x[i]-x[i-1])< epsilon then
       return((print("La aproximación buscada es
       x(",i,") = ",x[i])))
       )
       )$
```

Apliquemos dicha función al ejemplo anterior

```
(%i23) f(x):=%e^-x-x;newton(f,0.5,10^-8,20)$
```

(%o23) $f(x) := e^{-x} - x$
La aproximación buscada es $x(4) = 0{,}56714329040978$ (que es la cuarta aproximación obtenida).

En tanto que el error absoluto de esta aproximación será menor que 10^{-16}, pues resulta

```
(%i25) abs(f(x[4]))/1.3;
```

(%o25) $8{,}5401771125012034\,10^{-17}$

Veamos ahora que ocurre si ponemos $nmax = 3$

```
(%i26) f(x):=e^-x-x;newton(f,0.5,10^-8,3)$
```

(%o26) $f(x) := e^{-x} - x$

En este último caso al no lograrse la aproximación deseada en 3 pasos no da solución aproximada, si se desea se puede incrementar el número máximo de iteraciones “nmax”.

2.7. Problemas y trabajos propuestos

Problemas propuestos:

1. Utilizar el método de bisección para encontrar una solución aproximada, con error menor que 10^{-2} en $[4, 4{,}5]$, de la ecuación $x = \tan x$.

2. Probar que la ecuación $x - cosx = 0$ tiene una única raíz. Considerar el esquema iterativo: tomar x_0 arbitrario y calcular $x_{k+1} = cos(x_k)$ $(k = 0, 1, 2, \dots)$, probar su convergencia.

3. Utilizando el método iterativo del punto fijo, encontrar una raíz próxima a $x_0 = 0$ de la ecuación $2^x - 5x + 2 = 0$, trabajar con cuatro decimales redondeados e iterar hasta que $\mid x_i - x_{i-1} \mid < 0{,}00005$.

4. Realizar dos iteraciones con el método de Lagrange para aproximar la raíz de $2x^3 + x - 2 = 0$ en el intervalo $[0, 1]$, utilizando cuatro decimales redondeados en cada iteración.

5. Sean $m \in \mathbb{R}$ y $\mid \varepsilon \mid < 1$, probar que la ecuación $x = m - \varepsilon senx$ tiene una única solución en el intervalo $[m - \pi, m + \pi]$ y proponer un método iterativo para aproximarla.

6. Aplicar el método de bisección a la funcióm $f(x) = x^3 - 17$ para determinar la raíz cúbica de 17 con error menor que 10^{-2}, iniciar los cálculos en el intervalo $[2, 3]$. Asimismo, aproximarla utilizando un método iterativo adecuado y el método de Newton.

7. Para la ecuación $3x + senx - e^x = 0$, utilizando el método de Newton, aproximar la raíz próxima a $x_0 = 0$, redondeando los cálculos a cinco cifras significativas e iterando hasta que $\mid x_i - x_{i-1} \mid < 0{,}001$.

8. Acotar las raíces reales de la ecuación $x^4 + 2x^3 - 7x^2 + 3 = 0$, y determinar su número. Calcular con error menor que $5 \cdot 10^{-6}$, mediante el método de Newton, una de sus raćes positivas (justificando todos

los extremos del teorema que garantice la convergencia del método). Comprobar el orden de convergencia del método de Newton en este caso.

9. Probar que la ecuación $f(x) = e^x - 3x^2 = 0$ posee tres raíces reales distintas, aproximar la mayor de ellas utilizando el método de Newton, la menor mediante el método de Lagrange y la intermedia por un método convergente ideado por vosotros, realizando cuatro pasos y estimando el error correspondiente a la cuarta iteración; justificar los procedimientos utilizados.

10. Sea $f(x)$ una función de clase $\mathbb{C}^{(3)}$ con una única raíz simple en el intervalo $[c, d]$; determinar unas funciones $a(x)$ y $b(x)$ de modo que el método:

$$x_{n+1} = x_n + a(x_n)f(x_n) + b(x_n)f(x_n)^2 \ (n = 0, 1, 2, \dots)$$

sea al menos de tercer orden.

11. Dada la expresión

$$g(x) = x - \frac{f(x)}{f'(x)} + h(x)(\frac{f(x)}{f'(x)})^2$$

Se pide elegir $h(x)$ para que la iteración definida por $g(x)$ converja cúbicamente a una solución de $f(x) = 0$. Aplicar dicha iteración para aproximar la $\sqrt{10}$, partiendo del extremo de $[3, 4]$ para el que se puede asegurar la convergencia del método de Newton; realizar tres iteraciones.

12. Realizar al menos una iteración con el método de Newton para sistemas, a fin de aproximar la raíz del sistema

$$\begin{aligned} x - x^2 - y^2 &= 0 \\ y - x^2 + y^2 &= 0 \end{aligned}$$

próxima a $x_0 = 0{,}8$, $y_0 = 0{,}4$.

13. Con el método de Newton para sistemas, realizar dos iteraciones a fin de aproximar la raíz del sistema

$$\begin{aligned} x^2 - 10x + y^2 + 8 &= 0 \\ xy^2 + x - 10y + 8 &= 0 \end{aligned}$$

partiendo del punto $(0, 0)$.

Trabajos propuestos:

- Variantes del método de Newton-Raphson.
- Sucesiones de Sturm y aplicación a la separación de raíces reales, caso de las ecuaciones polinómicas.
- Métodos numéricos especiales para ecuaciones polinómicas: método de Bairstow.
- Ecuaciones en diferencias y método de Bernoulli para el cálculo de la raíz dominante de un polinomio.
- Variantes del método de Newton para sistemas no lineales.

Capítulo 3

Resolución numérica de sistemas lineales

3.1. Introducción. Normas vectoriales y matriciales

La teoría de ecuaciones lineales dio lugar al Álgebra Lineal y el estudio de su resolución práctica es de gran importancia, pues la mayor parte de los problemas en ciencias e ingeniería, una vez linealizados y discretizados conducen, en general, a la resolución de grandes sistemas de ecuaciones lineales. Conviene recordar a este respecto el teorema de Rouché-Frobenius sobre la existencia o no de soluciones de un sistema lineal. Consideraremos, salvo aviso en contra, sistemas compatibles determinados $Ax = b$ y veremos diversos métodos de cálculo de la única solución existente $x = A^{-1}b$. Los métodos de resolución de sistemas lineales los podemos clasificar en **directos** (que persiguen obtener la solución exacta en un número finito de pasos, como por ejemplo los métodos de Cramer, Gauss, LU, Cholesky, etc.) e **iterativos** (en los que se parte de unos valores aproximados iniciales y a partir de ellos se van obteniendo valores cada vez más aproximados de la solución x, ejemplos de tales métodos son los de Jacobi, Gauss-Siedel, relajación, etc.). Estos últimos son los que se suelen utilizar para aproximar las soluciones de sistemas lineales con un gran número de ecuaciones e incógnitas.

Para unos y otros se define la **estabilidad numérica** como la sensibilidad del método utilizado respecto a pequeñas perturbaciones o errores en los datos. Seguidamente damos las definiciones de normas vectoriales y matriciales que nos permitirán dar resultados sobre convergencia y acotación de errores en unos y otros.

Definición 2 *Sea V un espacio vectorial real o complejo, una* ***norma*** *sobre*

V es una aplicación $\| \cdot \|: V \to \mathbb{R}$ verificando las propiedades siguientes:

1. *$\forall x \in V$ es $\| x \| \geq 0$ y $\| x \| = 0 \Leftrightarrow x = 0$*
2. *$\forall x \in V$ y $\forall \alpha \in \mathbb{K}$ es $\| \alpha x \| = | \alpha | \| x \|$ ($\mathbb{K}$ es el cuerpo de los números reales o complejos).*
3. *$\forall x, y \in V$ es $\| x + y \| \leq \| x \| + \| y \|$*

Ejemplos de normas en $\mathbb{R}^n$ y $\mathbb{C}^n$ son $\| x \|_p = (\sum_{i=1}^n | x_i |^p)^{\frac{1}{p}}$ para $1 \leq p < \infty$ y $\| x \|_\infty = \max\{| x_i | \mid 1 \leq i \leq n\}$ (denominada norma del supremo); casos especialmente interesantes son los relativos a $p = 1$ y $p = 2$ dados por $\| x \|_1 = (\sum_{i=1}^n | x_i |)$ y $\| x \|_2 = (\sum_{i=1}^n | x_i |^2)^{\frac{1}{2}} = (\bar{x}^t x)^{1/2}$, aquí x denota la matriz columna con las mismas componentes que el vector x, en tanto que $\bar{x}^t$ es la matriz traspuesta de la conjugada de x, es decir la matriz fila cuyas componentes son las conjugadas de las de x (esta es denominada norma euclídea del vector x). Se puede probar que en los espacios vectoriales de dimensión finita sobre el cuerpo $\mathbb{K}$, de los números reales o complejos, todas las normas son equivalentes, en el sentido de que definen los mismos abiertos y cerrados con la topología métrica correspondiente a la distancia asociada a la norma definida por $d(x, y) = \| x - y \|$.

Definición 3 *Sobre el espacio vectorial de las matrices cuadradas de orden n reales o complejas $M_n(\mathbb{K})$, una norma se dice que es una **norma matricial** si además de las tres propiedades anteriores verifica la siguiente:*

- *$\forall A, B \in M_n(\mathbb{K})$ es $\| AB \| \leq \| A \| \| B \|$*

Definición 4 *Una norma vectorial $\| \cdot \|_V$ sobre $\mathbb{K}^n$ y una norma matricial $\| \cdot \|_M$ sobre $M_n(\mathbb{K})$ se dicen **compatibles** si se verifica la siguiente propiedad:*

- *$\forall A \in M_n(\mathbb{K})$ y $\forall x \in \mathbb{K}^n$ es $\| Ax \|_V \leq \| A \|_M \| x \|_V$*

El teorema siguiente, que enunciamos pero cuya demostración no haremos, nos dice como construir una norma matricial compatible con una norma vectorial dada.

Teorema 13 *Si $\| \cdot \|_V$ es una norma vectorial sobre $\mathbb{K}^n$ y $\forall A \in M_n(\mathbb{K})$ definimos $\| A \| = \max\{\| Ax \|_V$ siendo $\| x \|_V = 1\}$, entonces $\| \cdot \|$ es una norma matricial compatible con $\| \cdot \|_V$. Se denomina norma **inducida o subordinada** por $\| \cdot \|_V$.*

Observación.- También puede probarse sin dificultad que dada una norma matricial cualquiera $\| \cdot \|_M$ existen normas vectoriales $\| \cdot \|_V$ compatibles con ella, en efecto dado un vector $a \neq \overline{0}$ basta con definir para cualquier otro x, la norma $\| x \|_V = \| x a^t \|_M$.

3.1.1. Normas matriciales inducidas

Puede probarse, pero no lo haremos, que las normas matriciales inducidas por $\| \cdot \|_1$, $\| \cdot \|_2$ y $\| \cdot \|_\infty$ que indicamos por los mismos símbolos resultan ser:

- $\| A \|_1 = \max_j \sum_{i=1}^{n} | a_{ij} |$ (o sea es el máximo de las normas subuno de los vectores columna de la matriz dada).

- $\| A \|_2 = \rho(A^* A)^{\frac{1}{2}}$ (se llama norma espectral, donde $A^* = \overline{A}^t$ es la denominada matriz asociada de A que es la traspuesta de su matriz conjugada, si A fuese real su asociada sería simplemente su traspuesta). Conviene recordar que el **radio espectral** de una matriz M, $\rho(M) = \max_{1 \leq i \leq n} \{ | \lambda_i | \text{ tal que } \lambda_i \text{ es valor propio de } A \}$.

- $\| A \|_\infty = \max_i \sum_{j=1}^{n} | a_{ij} |$ (o sea es el máximo de las normas subuno de los vectores fila de la matriz dada).

3.1.2. Relación entre el radio espectral y la norma matricial de una matriz A

Conviene en este punto recordar el siguiente resultado técnico.

Teorema 14 *Para toda norma matricial $\| \cdot \|$ y toda matriz $A \in M_n(\mathbb{R})$ se verifica que $\rho(A) \leq \| A \|$. Además, el radio espectral de A está dado por*

$$\rho(A) = \inf\{ \| A \| : \| \cdot \| \text{ es una norma matricial} \}.$$

3.1.3. Número de condición de una matriz

Definición 5 *Con respecto a la resolución de sistemas lineales $Ax = b$, si la matriz A es inversible, se denomina **número de condición (o condicionamiento)** de dicha matriz respecto a una norma matricial dada al número*

$$k(A) = \|A\| \|A^{-1}\|$$

*Cuando A no es inversible se define $k(A) = \infty$. Es fácil probar que siempre se cumple $k(A) \geq 1$, cuando es $k(A) = 1$ se dice que A está **perfectamente condicionada**.*

Análisis de errores

Sea $Ax = b$ un sistema lineal compatible determinado con solución única $x = A^{-1}b$, si se perturba la matriz de términos independientes del sistema con $\triangle b$, la nueva solución será $x + \triangle x$, verificándose

$$A(x + \triangle x) = b + \triangle b$$

Entonces, se tendrá $A\triangle x = \triangle b$ o también $\triangle x = A^{-1}\triangle b$ y tomando cualquier norma vectorial y su norma matricial inducida resultará

$$\|\triangle x\| \leq \|A^{-1}\|\|\triangle b\|$$

lo que nos da una acotación del error absoluto $\triangle x$. Si se desea obtener una acotación del error relativo en x debido a la perturbación de los términos independientes $\triangle b$, basta con escribir

$$\|\triangle x\| \leq \|A^{-1}\|\|\triangle b\| = \|A^{-1}\|\|Ax\|\tfrac{\|\triangle b\|}{\|b\|} \leq \|A^{-1}\|\|A\|\|x\|\tfrac{\|\triangle b\|}{\|b\|}$$

resultando

$$\tfrac{\|\triangle x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\tfrac{\|\triangle b\|}{\|b\|} = k(A)\tfrac{\|\triangle b\|}{\|b\|}$$

es decir que el error relativo en x es menor o igual que $k(A)$ veces el error relativo en b.

Si el número de condición es grande se dice que la matriz está **mal condicionada** y la solución puede ser muy sensible con respecto a pequeños cambios en los términos independientes, en caso contrario se dice bien condicionada.

En general si se perturban tanto la matriz de coeficientes como los términos independientes con $\triangle A$ y $\triangle b$ resulta una solución $x + \triangle x$ que cumple

$$(A + \triangle A)(x + \triangle x) = b + \triangle b$$

y se puede probar el siguiente teorema.

Teorema 15 *Si* $\|\triangle A\| < \frac{1}{\|A^{-1}\|}$ *entonces, usando normas vectoriales y matriciales compatibles, se verifica*

$$\tfrac{\|\triangle x\|}{\|x\|} \leq \tfrac{k(A)\|I\|}{1-k(A)\frac{\|\triangle A\|}{\|A\|}}\left(\tfrac{\|\triangle A\|}{\|A\|} + \tfrac{\|\triangle b\|}{\|b\|}\right)$$

Que nos da una acotación del error relativo de la solución en estas condiciones.

Ejercicios. Se pide resolver los siguientes:

- Calcular las normas $\| \cdot \|_1$ y $\| \cdot \|_\infty$ de la matriz $\begin{pmatrix} -1 & 2 & 0 \\ 0 & -2 & -2 \\ 3 & 5 & 4 \end{pmatrix}$.

- Para una matriz ortogonal A, obtener $\| A \|_2$.

- Sea f la función de la forma $f(x) = Ax^{12} + Bx^{13}$, verificando que $f(0{,}1) = 6{,}06 \cdot 10^{-13}$ y $f(0{,}9) = 0{,}03577$. Determinar A y B, y evaluar la sensibilidad de estos parámetros respecto de pequeños cambios en los valores de f.

3.2. Métodos directos de resolución de sistemas lineales

Teorema 16 *En general, dado un sistema $Ax = b$ de m ecuaciones lineales con n incógnitas, si a la matriz ampliada $(A, b)_{m\times(n+1)}$ le efectuamos operaciones elementales de filas (lo cual equivale a multiplicar por matrices regulares de orden m a izquierda), se obtiene una nueva matriz ampliada $(\overline{A}, \overline{b})_{m\times(n+1)}$; entonces x_0 es solución del sistema $Ax = b$ si y sólo si es solución del sistema $\overline{A}x = \overline{b}$, es decir ambos sistemas son equivalentes.*

Demostración. En efecto, puesto que $(\overline{A}, \overline{b}) = M_p \cdots M_1(A, b) = (MA, Mb)$, siendo $M = M_p \cdots M_1$ una matriz regular por ser producto de un número finito de matrices regulares, se sigue que x_0 es solución de $Ax = b \Leftrightarrow Ax_0 = b \Leftrightarrow MAx_0 = Mb \Leftrightarrow \overline{A}x_0 = \overline{b} \Leftrightarrow x_0$ es solución de $\overline{A}x = \overline{b}$. Este método de resolución de sistemas de ecuaciones lineales se conoce como método de eliminación de incógnitas.

Observación. Dado que todo sistema de ecuaciones lineales con coeficientes complejos es equivalente a un sistema de doble número de ecuaciones e incógnitas con coeficientes reales, consideraremos en lo que sigue tan sólo sistemas lineales con coeficientes reales.

Seguidamente, veamos como resolver fácilmente sistemas triangulares y luego como llevar un sistema lineal dado a forma triangular.

3.2.1. Sistemas triangulares

Consideremos un sistema compatible determinado de la forma

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Es evidente que de la última ecuación resulta $x_n = b_n/a_{nn}$, este valor se puede sustituir en la penúltima ecuación y de ahí despejar x_{n-1} y reiterando el proceso se obtiene fácilmente la solución. El cálculo de cualquier incógnita viene dado por el algoritmo

$$\begin{array}{c} x_n = b_n/a_{nn} \\ x_i = (b_i - \sum_{j=i+1}^{n} a_{ij}x_j)/a_{ii} \ (i = n-1, \dots 2, 1) \end{array}$$

este cálculo es sencillo y rápido de realizar con ordenador. Es fácil ver que el costo computacional de esta resolución asciende a n^2 operaciones (entre sumas, productos y divisiones).

3.2.2. Eliminación gaussiana: el método de Gauss y sus variantes

El método de Gauss

Sea un sistema $Ax = b$ compatible determinado, es decir A es una matriz cuadrada de orden n con $|A| \neq 0$, el método de Gauss consiste en llevar el sistema a forma triangular superior, produciendo ceros en la matriz de coeficientes por debajo de la diagonal, realizando siempre operaciones elementales de filas. Recordemos que hay tres operaciones elementales de filas, a saber intercambiar dos filas entre sí (por ejemplo la fila i por la fila j), multiplicar una fila por un escalar no nulo (por ejemplo multiplicar todos los elementos de la fila i por el número α, siendo $\alpha \neq 0$) y añadir a una fila otra diferente multiplicada por un número cualquiera (por ejemplo sumar a la fila i la j multiplicada por un escalar β). Fácilmente se prueba que cada una de estas operaciones equivale a multiplicar por una matriz elemental a izquierda, así la primera equivale a multiplicar por la matriz de permutación P_{ij} obtenida al intercambiar las filas i y j de la matriz identidad, la segunda equivale a multiplicar a izquierda por la matriz diagonal $P_i(\alpha)$ con unos en la diagonal, salvo el elemento $a_{ii} = \alpha \neq 0$ y la tercera equivale a multiplicar a izquierda por la matriz $P_{ij}(\beta)$ cuyos elementos elementos de la diagonal son

iguales a 1 y los restantes nulos , salvo el $a_{ij} = \beta$, además es obvio que estas matrices son regulares o inversibles, luego su producto también lo es.

Para comenzar a describir el método, llamemos $A^{(1)} = A$ y $b^{(1)} = b$; así, partiendo de $A^{(1)}x = b^{(1)}$ realizamos las operaciones elementales de filas para ir obteniendo los sistemas equivalentes: $A^{(1)}x = b^{(1)}, \ldots, A^{(k)}x = b^{(k)}$ $(k = 1, 2, \ldots, n)$, donde $A^{(k)}$ tiene nulos los elementos por debajo de la diagonal en las $k-1$ primeras columnas, de esta forma obtendremos en $n-1$ pasos un sistema $A^{(n)}x = b^{(n)}$ donde $A^{(n)}$ es una matriz triangular superior, que se resolvería por el algoritmo indicado en la sección anterior.

Veamos con más precisión el primer paso a realizar para pasar del sistema $A^{(1)}x = b^{(1)}$ al $A^{(2)}x = b^{(2)}$, se pueden dar dos casos:

1. Si $a_{11}^{(1)} \neq 0$ a la fila i le sumamos la primera multiplicada por $-\frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$ y esto para cada $i = 2, 3, \ldots, n$.
2. Si $a_{11}^{(1)} = 0$, se busca la primera fila cuyo elemento de la primera columna sea distinto de cero y se intercambia con la primera fila mediante una permutación de filas, así estaríamos en el caso anterior y se procedería como se ha descrito en el apartado 1 anterior.

El paso general del sistema $A^{(k)}x = b^{(k)}$ al $A^{(k+1)}x = b^{(k+1)}$ se puede expresar como sigue:

1. Si $a_{kk}^{(k)} = 0$ se permuta esta fila con la primera fila posterior cuyo elemento de la columna k sea no nulo, se obtiene así una matriz equivalente

$$\widetilde{A}^{(k)} = P_{kt_k} A^{(k)} \text{ siendo } t_k = \min\{t \geq k \mid a_{tk}^{(k)} \neq 0\},$$
$$\text{también hacemos } \widetilde{b}^{(k)} = P_{kt_k} b^{(k)}$$

seguidamente, a las filas posteriores a la k le sumamos la fila k multiplicada por $-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ y esto para cada $i = k+1, \ldots, n$, todo lo cual equivale a hacer

$$A^{(k+1)} = M^{(k)} \widetilde{A}^{(k)} \text{ y } b^{(k+1)} = M^{(k)} \widetilde{b}^{(k)}$$

siendo

$$M^{(k)} = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & m_{k+1,k} & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & m_{n,k} & 0 & \ldots & 1 \end{pmatrix}$$

la denominada matriz de Frobenius, con unos en la diagonal, con elementos significativos en la columna k bajo la diagonal definidos por $m_{ik} = -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ $(i = k+1, \ldots, n)$ y siendo nulos los restantes elementos. El efecto de multiplicar por esta matriz es el de sacar ceros por debajo de la diagonal en la columna k.

2. Si $a_{kk}^{(k)} \neq 0$ entonces $t = k$ y $P_{kt_k} = I$ no siendo necesario el intercambio de filas y quedando $\widetilde{A}^{(k)} = A^{(k)}$ y $\widetilde{b}^{(k)} = b^{(k)}$.

El elemento $a_{kk}^{(k)}$ se llama pivote, y el hecho de figurar en el denominador puede acarrear errores de truncamiento grandes al tratar la ecuación por ordenador, por ello hace falta un método más estable de elegir dicho pivote en cada paso del método, las **estrategias de elección de pivote** más usuales son las siguientes:

a) **Pivote parcial**. Esta estrategia consiste en elegir la fila t_k a intercambiar con la k de manera que $\mid a_{t_k k}^{(k)} \mid = \text{máx}\{\mid a_{tk}^{(k)} \mid \mid t \geq k\}$, es decir consiste en elegir para intercambiar con la fila k no la primera fila posterior a la k con elemento de la columna k no nulo, sino la que tiene dicho elemento con módulo máximo, aún así puede ocurrir que, en algún caso, este método también de errores apreciables, por ello puede ser necesario usar la siguiente estrategia.

b) **Pivote total**. En este caso, se elige como pivote el elemento $\mid a_{t_k s_k}^{(k)} \mid = \text{máx}\{\mid a_{ts}^{(k)} \mid \mid k \leq t, s \leq n\}$, el problema en este caso es que pueden cambiarse las columnas y, por tanto, las incógnitas y esto ha de tenerse en cuenta.

Nota informativa. El método de Gauss puede llevarse a acabo de modo estable sin intercambio de filas para matrices estrictamente diagonal dominantes y para matrices simétricas definidas positivas.

Observación. Si dado el sistema $Ax = b$ con $\mid A \mid \neq 0$, puede aplicarse el método de Gauss sin permutaciones de filas, entonces se tiene que $A^{(n)} = M^{(n-1)} \cdot M^{(n-2)} \cdots M^{(1)} \cdot A$, pero la inversa de cada matriz de Frobenius es otra del mismo tipo con los elementos de la columna k por debajo de la diagonal opuestos de los de la matriz de partida, si despejamos la matriz A en la fórmula anterior, se tendrá $A = M^{(1)-1} \cdots M^{(n-1)-1} \cdot A^{(n)} = LU$, siendo $L = M^{(1)-1} \cdots M^{(n-1)-1}$ una matriz triangular inferior con unos en la diagonal y $U = A^{(n)}$ una matriz triangular superior con elementos diagonales

no nulos, pues $\mid A \mid = \mid U \mid$. En definitiva A se ha factorizado en la forma $A = LU$, donde L es una matriz triangular inferior y U triangular superior. En todo caso si $\mid A \mid \neq 0$ hay una reordenación de las filas de A tal que la nueva matriz es factorizable de esta forma.

Algoritmo y costo computacional del método de Gauss. Dar el algoritmo para resolver un sistema triangular superior, así como del método de Gauss para llevar a forma triangular un sistema compatible determinado (en el caso de no ser necesarias permutaciones de filas). Asimismo, probar que para resolver completamente dicho sistema por el método de Gauss son necesarias $(2n^3 + 3n^2 - 5n)/6$ sumas, un número igual de multiplicaciones y $(n^2 + n)/2$ divisiones; en definitiva, en este caso el costo computacional del método de Gauss para n grande es del orden de $\frac{2}{3}n^3$.

Ejercicio. Supongamos que el determinante se calcula directamente con su definición, como suma de todos los productos posibles de n elementos de manera que haya uno de cada fila y columna afectados del signo positivo o negativo según que las permutaciones de los índices de fila y columna correspondientes sean de la misma o distinta paridad, y que resolvemos un sistema de n ecuaciones lineales con n incógnitas por el método de Cramer y por el método de Gauss, si en cada operación elemental es necesario un tiempo de $1{,}5 \cdot 10^{-6}$ segundos, estimar el tiempo necesario para resolver un sistema de $n = 5, 10, 20$ ecuaciones por el método de Cramer y por el método de Gauss.

Método de Gauss-Jornan

Consiste en llevar la matriz a forma diagonal, en vez de eliminar x_k solamente de las filas i con $i > k$, se elimina de todas las filas $i \neq k$, con lo cual la matriz del sistema queda diagonal despejándose fácilmente las variables. A su vez, puede aplicarse con estrategia de pivote parcial o total. El costo computacional se incrementa con respecto al método de Gauss.

3.2.3. Otros métodos de factorización

Factorización LU

Si dado un sistema lineal $Ax = b$ con $\mid A \mid \neq 0$, se conoce una descomposición de A en la forma $A = LU$ con L triangular inferior y U triangular superior, la resolución de dicho sistema se reduce a la resolución de dos sistemas triangulares, pues llamando $y = Ux$, bastaría resolver primero el sistema

$Ly = b$ y a continuación el $Ux = y$, con el costo de $2n^2$ operaciones. El problema sería saber si dicha factorización existe y cómo calcularla. Veamos a continuación respuestas a estas cuestiones.

Teorema 17 *Sea $A \in M_n(\mathbb{R})$ tal que todos los determinantes de las submatrices principales $A_k = (a_{ij})_{1\leq i,j\leq k}$ $(k = 1, 2, \ldots, n)$ son distintos de cero. Entonces, A es descomponible como producto de una matriz L, triangular inferior con $l_{ii} = 1$ para todo $i = 1, 2, \ldots, n$, por U triangular superior. Siendo única la descomposición en las condiciones anteriores.*

Demostración. Puede hacerse por inducción pero la omitimos.

Cálculo directo de la factorización. En las hipótesis del teorema anterior la factorización LU de la matriz A es posible, calculándose los l_{ij} sin utilizar el método de Gauss, pues de $A = LU$ resulta

$$a_{ij} = \sum_{p=1}^{h} l_{ip}u_{pj} \text{ siendo } h = \text{mín}\{i, j\}$$

o sea

$$a_{kj} = l_{k1}u_{1j} + l_{k2}u_{2j} + \cdots + l_{kk}u_{kj} \text{ (para } j \geq k) \text{ (i)}$$
$$a_{ik} = l_{i1}u_{1k} + l_{i2}u_{2k} + \cdots + l_{ik}u_{kk} \text{ (para } i \geq k) \text{ (i')}$$

y hemos tomado $l_{kk} = 1$ para todo k. Una vez descompuesta A hay que resolver los dos sistemas triangulares

$$Ly = b,\ Ux = y \text{ (ii)}$$

el primero de los cuales se resuelve mediante:

$$b_k = l_{k1}y_1 + l_{k2}y_2 + \cdots + l_{kk}y_k \ (k = 1, 2, \ldots, n) \text{ (iii)}$$

relación análoga a la (i). En consecuencia el algoritmo:

- Para cada $k = 1, 2, \ldots, n$, hacer:

  - Para $j = k, k + 1, \ldots, n$, calcular:

  $$u_{kj} = a_{kj} - \sum_{p=1}^{k-1} l_{kp}u_{pj}$$

  - $y_k = b_k - \sum_{p=1}^{k-1} l_{kp}y_p$

- Para $i = k+1, k+2, \ldots, n$, calcular:

$$l_{ik} = (a_{ik} - \sum_{p=1}^{k-1} l_{ip} u_{pk})/u_{kk}$$

este algoritmo realiza la factorización a la vez que calcula $y = L^{-1}b$, luego habría que resolver el sistema triangular $Ux = y$, en conjunto se conoce como **método de Doolitle-Banachiewicz**, si se tomasen unos en la diagonal de U, tendríamos el denominado **método de Crout**.

Factorización de Cholesky

Teorema 18 *Para matrices A reales son equivalentes:*

I) *A es simétrica y definida positiva*

II) *Existe L triangular inferior real, con elementos diagonales positivos, tal que $A = LL^t$.*

Demostración. Se hace por inducción pero la omitimos.

Cálculo directo de la factorización de Cholesky. En las hipótesis del teorema anterior es $A = LL^t$, por tanto será

$$a_{ij} = \sum_{k=1}^{n} l_{ik} l_{kj}^t = \sum_{k=1}^{h} l_{ik} l_{jk} \text{ siendo } h = \min\{i, j\}$$

ahora para $i = j$ resulta

$$a_{jj} = \sum_{k=1}^{j} l_{jk}^2 = \sum_{k=1}^{j-1} l_{jk}^2 + l_{jj}^2$$

de donde se deduce que

$$\boxed{l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}}$$

en tanto que para $i = j+1, j+2, \ldots, n$ se tendrá

$$a_{ij} = \sum_{k=1}^{j} l_{ik} l_{jk} = \sum_{k=1}^{j-1} l_{ik} l_{jk} + l_{ij} l_{jj}$$

de la que se obtienen los

$$l_{ij} = (a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk})/l_{jj}$$

Una vez hallada la matriz L mediante el algoritmo descrito, se resolverían los dos sistemas triangulares $Ly = b$ y $L^t x = y$, con un costo computacional total para n grande del orden de $n^3/3$.

Ejercicios.

1. Resolver, según el método de Doolitle, el sistema $Ax = (1, 1, -1)^t$ siendo la matriz $A = \begin{pmatrix} 1 & 1 & 2 \\ -1 & 0 & 1 \\ 2 & 1 & -1 \end{pmatrix}$.

2. Resolver aplicando el método de Cholesky, el sistema $Ax = (1, 1, 0)^t$ siendo la matriz $A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 5 & 1 \\ 1 & 1 & 3 \end{pmatrix}$.

3.3. Métodos iterativos de resolución de sistemas lineales

Definición 6 *Una sucesión de matrices cuadradas, de orden n,* $\{A^{(k)}\}_1^{\infty}$ *tiene por límite la matriz A del mismo orden si* $\lim_{k\to\infty} \| A^{(k)} - A \| = 0$ *para alguna norma matricial. Y se dice que una matriz A es* ***casinilpotente*** *si* $\lim_{m\to\infty} A^m = 0$.

Nos será útil el siguiente teorema, cuya demostración omitiremos.

Teorema 19 *Una matriz A, de orden n, es casinilpotente* $\Leftrightarrow \rho(A) < 1$.

3.3.1. Generalidades: convergencia y construcción de métodos iterativos

Consideraremos aquí métodos iterativos de resolución de sistemas lineales $Ax = b$ con $|A| \neq 0$, en los que se parte de una aproximación inicial de la solución $x^{(0)}$ del sistema y se generan las aproximaciones sucesivas $\{x^{(k)}\}_0^{\infty}$ en la forma

$$\boxed{x^{(k+1)} = Ex^{(k)} + F \ (k = 0, 1, 2, \dots)} \quad (1)$$

donde $E \in M_n(\mathbb{R})$ se denomina **matriz de iteración** y $F \in M_{n\times 1}(\mathbb{R})$.

Definición 7 *Un método (1) como el descrito se dice* ***convergente*** *para el sistema* $Ax = b$ *si* $\forall x^{(0)} \in \mathbb{R}^n : \lim_{k\to\infty} x^{(k)} = x = A^{-1}b$.

Desde luego si el método es convergente, tomando límites en (1), la solución ha de verificar la relación

$$x = Ex + F \ (2)$$

o lo que es equivalente

$$F = (I - E)A^{-1}b \ (3)$$

aunque puede no verificarse el recíproco, por ello se introduce la siguiente definición.

Definición 8 *Un método* (1) *se dice* ***consistente con el sistema*** $Ax = b$ *si se verifica la condición* (2) *o equivalentemente* (3).

A continuación enunciamos el **teorema fundamental** sobre convergencia de métodos iterativos de resolución aproximada de sistemas lineales (su demostración puede verse en [7]).

Teorema 20 *El método* (1) *es convergente con respecto al sistema* $Ax = b$ *si y sólo si se verifican simultáneamente las dos condiciones siguientes:*

1. *El método es consistente con el sistema. Y*

2. $\rho(E) < 1$.

Construcción de métodos iterativos

Sea el sistema $Ax = b$ con A inversible cuya solución es $x = A^{-1}b$, descomponiendo A en la forma: $A = M - N$ con M inversible, entonces

$$\begin{gathered} Ax = b \Leftrightarrow (M - N)x = Mx - Nx = b \Leftrightarrow \\ \Leftrightarrow Mx = Nx + b \Leftrightarrow x = M^{-1}Nx + M^{-1}b \end{gathered}$$

expresión que sugiere el método

$$\boxed{x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b} \ (4)$$

los métodos así construidos son consistentes por tanto, según el teorema fundamental, serán convergentes si y sólo si $\rho(M^{-1}N) < 1$. Además, el teorema siguiente asegura que todo método iterativo convergente puede construirse de esa manera.

Teorema 21 *Todo método* (1) *convergente es de la forma* (4) *con* $A = M - N$ *y* M *inversible.*

Cota de error en los métodos iterativos convergentes

Sea $x^{(k+1)} = Ex^{(k)} + F$ un método iterativo convergente para $Ax = b$, denotando por $\| \cdot \|$ una norma matricial tal que $\| E \| < 1$ y una norma vectorial compatible con ella, se tendrá

$$\| x^{(m-1)} - x \| \leq \| x^{(m-1)} - x^{(m)} \| + \| x^{(m)} - x \| \ (*)$$

como además

$$x^{(m)} - x = Ex^{(m-1)} + F - (Ex + F) = E(x^{(m-1)} - x)$$

resultará que

$$\| x^{(m)} - x \| \leq \| E \| \| x^{(m-1)} - x \| \ (**)$$

Ahora, desde $(*)$ y $(**)$ es fácil obtener

$$(1 - \| E \|) \| x^{(m-1)} - x \| \leq \| x^{(m)} - x^{(m-1)} \| \ (***)$$

y de las desigualdades $(**)$ y $(***)$ se deduce

$$\boxed{\| x^{(m)} - x \| \leq \frac{\| E \|}{1 - \| E \|} \| x^{(m)} - x^{(m-1)} \| \leq \frac{\| E \|^m}{1 - \| E \|} \| x^{(1)} - x^{(0)} \|}$$

que puede utilizarse para estimar el error y como test de parada del programa correspondiente.

3.3.2. Métodos iterativos particulares: Jacobi, Gauss-Seidel y relajación

La expresión (4) anterior puede escribirse también en la forma

$$Mx^{(k+1)} = Nx^{(k)} + b \ (5)$$

donde $A = M - N$ y M inversible. Así pues, dado el sistema $Ax = b$, descompongamos la matriz $A = D - L - U$ con

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ 0 & 0 & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{pmatrix}$$

$$L = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ -a_{21} & 0 & 0 & \dots & 0 \\ -a_{31} & -a_{32} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & -a_{n3} & \dots & 0 \end{pmatrix}$$

$$U = \begin{pmatrix} 0 & -a_{12} & -a_{13} & \dots & -a_{1n} \\ 0 & 0 & -a_{23} & \dots & -a_{2n} \\ 0 & 0 & 0 & \dots & -a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

Método de Jacobi

En estas condiciones, el **método de Jacobi** consiste en hacer

$$\boxed{M = D \text{ y } N = L + U}$$

como M debe ser inversible es necesario que todo $a_{ii} \neq 0$ para $i = 1, 2, \dots, n$, con lo que puede expresarse en la forma

$$\boxed{Dx^{(k+1)} = (L + U)x^{(k)} + b}$$

y en componentes resulta

$$\boxed{x_i^{(k+1)} = (-\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)} + b_i)/a_{ii}}$$

se dice que es un método de **aproximaciones simultáneas**, pues para hallar un $x_i^{(k+1)}$ hace falta conocer todos los $x_i^{(k)}$, los cuales se utilizan simultáneamente. Este método será convergente si y sólo si $\rho(D^{-1}(L + U)) < 1$.

Método de Gauss-Seidel

El **método de Gauss-Seidel** consiste en tomar

$$\boxed{M = D - L \text{ y } N = U}$$

con lo cual puede expresarse en forma vectorial como

$$\boxed{(D - L)x^{(k+1)} = Ux^{(k)} + b}$$

y en componentes se expresa en la forma

$$\boxed{x_i^{(k+1)} = (-\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i)/a_{ii}}$$

se dice que es un método de **aproximaciones sucesivas**, pues se van usando las aproximaciones obtenidas en la pasada que se está realizando a partir del momento en que se calculan. Este método será convergente si y sólo si $\rho((D-L)^{-1}U) < 1$, se espera que sea más rápidamente convergente y así ocurre en muchos casos, pero este comportamiento no es general. Veamos dos resultados sobre convergencia.

Teorema 22 *Si la matriz de coeficientes del sistema $Ax = b$ es estrictamente diagonal dominante los métodos de Jacobi y Gauss-Seidel son convergentes, siendo los radios espectrales de las matrices de iteración correspondientes menores o iguales a c con*

$$\boxed{c = \max_{1\le i\le n}\{\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|\, a_{ij}\,|}{|\, a_{ii}\,|}\} < 1}$$

Teorema 23 *Para matrices tridiagonales, los radios espectrales de las matrices de iteración de Jacobi (E_J) y Gaus-Seidel (E_G) verifican la relación*

$$\boxed{\rho(E_G) = \rho(E_J)^2}$$

y por tanto ambos son simultáneamente convergentes o divergentes, y cuando ambos convergen el de Gauss-Seidel es asintóticamente más rápido que el de Jacobi.

Observación. Recordemos que una matriz se dice **estrictamente diagonal dominante** si para cada fila i se verifica que

$$|\, a_{ii}\,| > \sum_{\substack{j=1\\ j\neq i}}^{n} |\, a_{ij}\,|$$

Asimismo, una matriz se dice **tridiagonal** si $a_{ij} = 0$ para todo i, j tal que $|\, i-j\,| > 1$, es decir los elementos que no están en la diagonal principal o en una paralela por encima o por debajo son nulos.

Método de relajación

Los métodos del tipo de Gauss-Seidel se conocen con el nombre de métodos de relajación. Si en la descomposición de la matriz $A = D - L - U$, para $\omega \neq 0$ escribimos $D = \frac{1}{\omega}D - \frac{1-\omega}{\omega}D$, entonces nos quedará $A = \frac{1}{\omega}D - L - \frac{1-\omega}{\omega}D - U$ y tomando

$$\boxed{M = \frac{1}{\omega}D - L \text{ y } N = \frac{1-\omega}{\omega}D + U}$$

obtenemos el **método de relajación**, que se expresa en forma vectorial como

$$\boxed{(\frac{1}{\omega}D - L)x^{(k+1)} = (\frac{1-\omega}{\omega}D + U)x^{(k)} + b}$$

y en componentes por la fórmula

$$\boxed{\frac{1}{\omega}a_{ii}x_i^{(k+1)} + \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} = \frac{1-\omega}{\omega}a_{ii}x_i^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i}$$

De la que se despeja, fácilmente, $x_i^{(k+1)}$. Para cada valor de ω obtenemos un método distinto, en particular para $\omega = 1$ se obtiene el método de Gauss-seidel. Puede probarse que una condición necesaria de convergencia es que $0 < \omega < 2$. También puede probarse el siguiente.

Teorema 24 *Si A es estrictamente diagonal dominante y $0 < \omega \leq 1$ el método de relajación es convergente.*

3.4. Introducción al cálculo aproximado de valores y vectores propios

Otro problema importante en la práctica es el de determinar los valores y vectores propios de una matriz cuadrada A, de orden n. Recordemos que los valores propios de una matriz A son las raíces λ_i de la ecuación algebraica $| A - \lambda I |= 0$, donde I es la matriz identidad de orden n, en tanto que un vector u_i no nulo es un vector propio correspondiente al valor propio λ_i si $Au_i = \lambda_i u_i$. En general, el cálculo de los valores propios de A requeriría primero determinar el polinomio característico para luego intentar aproximar sus raíces, lo cual es a menudo imposible. Pero, en muchos problemas, sólo interesa el valor propio dominante es decir el de módulo máximo de la matriz A, supuesto existente. Veremos, en primer lugar un método de localización de

valores propios de una matriz A y luego se esbozará el método de potencias de cálculo aproximado de un valor propio dominante y un vector propio asociado. Luego, se podría reducir la matriz a otra de orden $n-1$ y se volvería a aplicar el método en un proceso llamado de deflación.

Teorema 25 ***(Teorema de Gerschgorin)*** *Sea A una matriz cuadrada de orden n y sean*

$$r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} | a_{ij} | \quad y \ c_j = \sum_{\substack{i=1 \\ i\neq j}}^{n} | a_{ij} |$$

entonces se tiene:

a) Todo valor propio de A está en la unión de los discos R_i, $(i = 1, 2, \dots, n)$ con $R_i = \{z \mid | z - a_{ii} | \leq r_i\}$

b) Todo valor propio de A está en la unión de los discos C_j, $(j = 1, 2, \dots, n)$ con $C_j = \{z \mid | z - a_{jj} | \leq c_j\}$

Método de potencias. Dada una matriz real de orden n, diagonalizable con un valor propio dominante, es decir verificando $| \lambda_1 | > | \lambda_2 | \geq \cdots \geq | \lambda_n |$, entonces λ_1 es real y por tanto puede tomarse un vector propio correspondiente u_1 de componentes reales. Además, si $x^{(0)}$ es un vector arbitrario y definimos $x^{(m+1)} = Ax^{(m)}$, la dirección de $x^{(m)}$ tiende, en general, a ser la de u_1 para $m \to \infty$, sucesión que conviene normalizar para evitar fuertes cambios de un paso al siguiente. Por otro lado, si dividimos las componentes k-ésimas de $x^{(m+1)}$ y $x^{(m)}$ la sucesión $\{\alpha_m\}_{m=1}^{\infty}$, definida por $\alpha_{m+1} = \frac{x_k^{(m+1)}}{x_k^{(m)}}$, converge al valor propio dominante λ_1.

3.5. Problemas resueltos

1. Averiguar si es posible la factorización $A = LU$ y calcularla en caso afirmativo, siendo la matriz

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Solución. Es fácil comprobar que los determinantes principales de A son todos no nulos, en efecto:

$$det(A_1) = |1| = 1 \neq 0, \ \det(A_2) = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1 \neq 0$$

y

$$\det(A_3) = \begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{vmatrix} = 2 \neq 0$$

entonces, en virtud del teorema correspondiente la matriz A admite la factorización LU con unos en la diagonal de L

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \cdot \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

El calculo de los elementos de L y U se hace ordenadamente como sigue, en primer lugar a partir de la primera fila de A, se calcula la primera fila de U mediante las ecuaciones:

$1 = 1 \cdot u_{11}$, $0 = 1 \cdot u_{12}$ y $1 = 1 \cdot u_{13}$ de donde se obtienen $u_{11} = 1$, $u_{12} = 0$ y $u_{13} = 1$. Seguidamente se igualan ordenadamente los elementos de la primera columna de A por debajo de la diagonal con los correspondientes del producto, obteniéndose las ecuaciones:

$0 = l_{21} \cdot u_{11} = l_{21} \cdot 1$, $-1 = l_{31} \cdot u_{11} = l_{31} \cdot 1$, de las que deducimos los elementos de la primera columna de L por debajo de la diagonal, que resultan ser $l_{21} = 0$ y $l_{31} = -1$.

Ahora se igualan los elementos de la segunda fila de A desde la diagonal hasta el final de dicha fila con los elementos del producto, que nos dan las ecuaciones:

$1 = l_{21} \cdot u_{12} + u_{22} = u_{22}$, $0 = l_{21} \cdot u_{13} + 1 \cdot u_{23} = u_{23}$, de las que se deduce la segunda fila de U, que resulta ser $u_{22} = 1$ y $u_{23} = 0$. Se procede igual con los de la segunda columna de A por debajo de la diagonal, que nos da la ecuación $0 = l_{31} \cdot u_{12} + l_{32} \cdot u_{22} = -1 \cdot 0 + l_{32} \cdot 1 = l_{32}$, de la que se obtiene $l_{32} = 0$. Finalmente, se iguala el elemento a_{33} con el producto de la tercera fila de L por la tercera columna de U, dando la ecuación $1 = l_{31} \cdot u_{13} + l_{32} \cdot u_{23} + 1 \cdot u_{33} = -1 \cdot 1 + 0 \cdot 0 + u_{33}$, de la que se deduce que $u_{33} = 2$. Se tiene pues la descomposición

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

2. Dada la matriz $A = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$, ¿cuál de las siguientes afirmaciones es correcta?:

a) Se puede factorizar en la forma $A = LL^t$ siendo $L = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$

b) La matriz A no admite la factorización de Cholesky, pues es simétrica pero no definida positiva

c) Las restantes afirmaciones son falsas
(Justificar la respuesta)

Solución. La afirmación a) es falsa, como se ve haciendo el producto LL^t, que siendo $L = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$ nos da $LL^t = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \neq A$.

La afirmación b) también es falsa, pues A es evidentemente simétrica y definida positiva, ya que sus menores principales

$$det(A_1) = |1| = 1 > 0, \ \det(A_2) = \begin{vmatrix} 1 & 2 \\ 2 & 5 \end{vmatrix} = 1 > 0$$

son positivos y por el teorema admite la factorización de Cholesky.

Por lo tanto ha quedado probada, que la afirmación c) es la correcta.

3. Utilizando el método de Cholesky, resolver el sistema

$$\begin{aligned} x_1 - x_2 + x_3 &= 1 \\ -x_1 + 5x_2 + x_3 &= 1 \\ x_1 + x_2 + 3x_3 &= 0 \end{aligned}$$

Solución. Puede comprobarse que la matriz de coeficientes A es simétrica y definida positiva (pues sus determinantes principales son positivos), por tanto admite la factorización de Cholesky $A = LL^t$, siendo L triangular inferior con elementos diagonales $l_{ii} > 0$, que se puede calcular a partir de la igualdad

$$\begin{pmatrix} 1 & -1 & 1 \\ -1 & 5 & 1 \\ 1 & 1 & 3 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \cdot \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix}$$

Debido a la simetría es suficiente con trabajar por filas; así, igualando los elementos correspondientes de la primera fila de A con los productos del segundo miembro, se obtienen las ecuaciones: $1 = l_{11}^2$, $-1 = l_{21}l_{11}$ y $1 = l_{31}l_{11}$; de las que se deduce que $l_{11} = 1$, $l_{21} = -1$ y $l_{31} = 1$. Seguidamente, se igualan los elementos de la segunda fila de A desde la diagonal hasta el final de dicha fila con los productos correspondientes

del segundo miembro, dando lugar a las ecuaciones: $5 = l_{21}^2 + l_{22}^2 = 1 + l_{22}^2$ y $1 = l_{21}l_{31} + l_{22}l_{32} = -1 + l_{22}l_{32}$, de las que deducimos que $l_{22} = 2$ y $l_{32} = 1$. Finalmente, igualando el elemento de la tercera fila y tercera columna de A con el producto correspondiente se obtiene la ecuación $3 = l_{31}^2 + l_{32}^2 + l_{32}^2 = 1 + 1 + l_{32}^2$ de la que se deduce que $l_{33} = 1$. Luego

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Resolver el sistema $Ax = b$ o sea $LL^tx = b$ es equivalente a resolver los sistemas $Ly = b$ y $L^tx = y$, que resultan ser

$$\begin{aligned} y_1 &= 1 \\ -y_1 + 2y_2 &= 1 \\ y_1 + y_2 + y_3 &= 0 \end{aligned}$$

cuya solución se obtiene por un proceso, denominado de descenso, obteniéndose $y_1 = 1$, $y_2 = 1$ e $y_3 = -2$; y

$$\begin{aligned} x_1 - x_2 + x_3 &= 1 \\ 2x_2 + x_3 &= 1 \\ x_3 &= -2 \end{aligned}$$

cuya solución se obtiene por un proceso denominado de remonte y nos da $x_3 = -2$, $x_2 = \frac{3}{2}$ y $x_1 = \frac{9}{2}$, y es la solución de problema planteado.

4. Probar que para el sistema: $ax + by = p$, $cx + dy = q$ $(ad \neq 0)$ una condición necesaria y suficiente de convergencia de los métodos iterativos de Jacobi y Gauss-Seidel es que

$$| bc | < | ad |$$

(Ayuda: Obtener primero las matrices de iteración de ambos métodos y luego sus valores propios).

Solución. Como sabemos los métodos de Jacobi y Gauss-Seidel son consistentes por construcción, luego convergen si y sólo si los radios espectrales de sus respectivas matrices de iteración son menores que 1. Calculemos pues las matrices de iteración de dichos métodos

- Para el método de Jacobi dicha matriz está dada por

$$E_J = D^{-1}(L + U)$$

con lo que resulta

$$E_J = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & -b \\ -c & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{b}{a} \\ -\frac{c}{d} & 0 \end{pmatrix}$$

cuyos valores propios son $\{-\sqrt{\frac{bc}{ad}}, \sqrt{\frac{bc}{ad}}\}$ si la fracción $\frac{bc}{ad} \geq 0$, o bien $\{-|\frac{bc}{ad}|i, |\frac{bc}{ad}|i\}$ en caso contrario, pero en cualquier caso será

$$\rho(E_J) = \sqrt{\mid \frac{bc}{ad} \mid} < 1 \Leftrightarrow \mid bc \mid < \mid ac \mid$$

como queríamos probar.

- En tanto que para el método de Gauss-Seidel la matriz de iteración es

$$E_{GS} = (D-L)^{-1}U$$

y calculando se tiene

$$E_{GS} = \begin{pmatrix} a & 0 \\ c & d \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & -b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{b}{a} \\ 0 & \frac{bc}{ad} \end{pmatrix}$$

cuyos valores propios son ahora $\{\frac{bc}{ad}, 0\}$, luego

$$\rho(E_{GS}) = \sqrt{\mid \frac{bc}{ad} \mid} < 1 \Leftrightarrow \mid bc \mid < \mid ac \mid$$

como queríamos probar.

5. Probar que si λ es un valor propio de la matriz cuadrada A entonces, para toda norma matricial $\parallel \parallel$, se verifica que $\mid \lambda \mid \leq \parallel A \parallel$, deducir de ello que

$$\rho(A) \leq \parallel A \parallel$$

Para $A = \begin{pmatrix} -1 & 0 & 1 \\ 2 & 0 & 1 \\ -1 & 0 & -2 \end{pmatrix}$, calcular $\rho(A)$, $\parallel A \parallel_2$ y comprobar en este caso la desigualdad anterior.

Solución. Si λ es un valor propio de A existe un vector x no nulo tal que $Ax = \lambda x$, y para toda norma matricial $\parallel \cdot \parallel$ existe una norma vectorial compatible con ella, que indicamos también por $\parallel \cdot \parallel$, tal que $\parallel Ax \parallel = \mid \lambda \mid \parallel x \parallel$, ahora bien por ser las normas matricial y vectorial compatibles se tiene que $\parallel Ax \parallel \leq \parallel A \parallel \parallel x \parallel$, pero de $\parallel \lambda x \parallel = \mid \lambda \mid \parallel x \parallel$,

por tanto $| \lambda | \| x \| \leq \| A \| \| x \|$, como $\| x \| \neq 0$ por ser x un vector no nulo, se deduce que $| \lambda | \leq \| A \|$ y siendo el radio espectral de A el máximo de los módulos de su valores propios, concluimos que $\rho(A) \leq \| A \|$, como queríamos demostrar.

Para la segunda parte, resolvemos la ecuación característica de la matriz A, que resulta ser:

$$| A - \lambda I | = -\lambda \left(\lambda^2 + 3\,\lambda + 3 \right) = 0$$

y obtenemos sus raíces características

$$\{-\frac{\sqrt{3}\,i + 3}{2}, \frac{\sqrt{3}\,i - 3}{2}, 0\}$$

por tanto $\rho(A) = \text{máx}\{| \frac{\sqrt{3}i+3}{2} |, | \frac{\sqrt{3}i-3}{2} |, 0\} = \text{máx}\{\sqrt{\frac{3}{4} + \frac{9}{4}}, 0\} = \sqrt{3}$.

Por otro lado, para calcular $\| A \|_2 = \sqrt{\rho(A^t A)}$, hacemos el producto

$$A^t A = \begin{pmatrix} -1 & 2 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & -2 \end{pmatrix} \cdot \begin{pmatrix} -1 & 0 & 1 \\ 2 & 0 & 1 \\ -1 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 6 & 0 & 3 \\ 0 & 0 & 0 \\ 3 & 0 & 6 \end{pmatrix}$$

Ahora la ecuación característica de $A^t A$ es

$$| A^t A - \lambda I | = -\lambda \left(\lambda^2 - 12\,\lambda + 27 \right) = 0$$

y sus valores propios son $\{3, 9, 0\}$, por tanto $\rho(A^t A) = 9$ y $\| A \|_2 = 3$, con lo que queda comprobada en este caso la desigualdad

$$\rho(A) = \sqrt{3} \leq \| A \|_2 = 3$$

6. Sea el sistema lineal $Ax = B$ con $| A | \neq 0$, probar que el método iterativo dado por

$$x^{(k+1)} = D^{-1}(D - \alpha A)x^{(k)} + \alpha D^{-1} B$$

con α real y D la diagonal de A, que se supone inversible (es decir con $a_{ii} \neq 0$ para todo $i \in \{1, 2, \ldots, n\}$) es consistente y dar una condición necesaria y suficiente para que sea convergente.

Solución. Recordemos que un método iterativo de la forma $x^{(k+1)} = Ex^{(k)} + F$ se dice **consistente** para el sistema $Ax = B$ si su solución, que podemos escribir en la forma $x = A^{-1}B$, verifica la ecuación de método, es decir debe ser $x = Ex + F$, veámoslo en nuestro caso $Ex +$

$F = D^{-1}(D - \alpha A)x + \alpha D^{-1}B = Ix - \alpha D^{-1}Ax + \alpha D^{-1}B = x - \alpha D^{-1}B + \alpha D^{-1}B = x$, luego el método propuesto es consistente. Como además es convergente para el sistema $Ax = B$ si y sólo si es consistente y radio espectral de la matriz de iteración ($E = D^{-1}(D - \alpha A)$) es menor que 1; concluimos que en este caso será convergente si y sólo si $\rho(D^{-1}(D - \alpha A)) < 1$.

7. Sea

$$A = \begin{pmatrix} 10 & -1 & 0 \\ -1 & 20 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

calcular $\| A \|_1$, $\| A \|_\infty$, $\| A \|_2$ y estudiar para que valores de α sería convergente el método iterativo descrito en el ejercicio anterior.

Solución. Recordemos que $\| A \|_1 = \text{máx}_j \sum_{i=1}^n | a_{ij} | = \text{máx}\{11, 21, 3\} = 21$, $\| A \|_\infty = \text{máx}_i \sum_{j=1}^n | a_{ij} | = \text{máx}\{11, 21, 3\} = 21$ y $\| A \|_2 = \rho(A^t A)^{\frac{1}{2}}$ ahora

$$A^t A = \begin{pmatrix} 10 & -1 & 0 \\ -1 & 20 & 0 \\ 0 & 0 & 3 \end{pmatrix} \cdot \begin{pmatrix} 10 & -1 & 0 \\ -1 & 20 & 0 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 101 & -30 & 0 \\ -30 & 401 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

la ecuación característica de $A^t A$ es

$$-(\lambda - 9)\left(\lambda^2 - 502\,\lambda + 39601\right) = 0$$

y sus valores propios son

$$\{251 - 30\sqrt{26}, 30\sqrt{26} + 251, 9\}$$

por tanto $\rho(A^t A) = 30\sqrt{26} + 251$ y $\| A \|_2 = \sqrt{30\sqrt{26} + 251} = 20{,}09901951359279$.

Por otro lado, el método del ejercicio anterior será convergente para todo α tal que $\rho(D^{-1}(D - \alpha A)) < 1$, ahora

$$D^{-1}(D - \alpha A) = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 3 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 10(1-\alpha) & \alpha & 0 \\ \alpha & 20(1-\alpha) & 0 \\ 0 & 0 & 3(1-\alpha) \end{pmatrix}$$

$$= \begin{pmatrix} 1-\alpha & \frac{\alpha}{10} & 0 \\ \frac{\alpha}{20} & 1-\alpha & 0 \\ 0 & 0 & 1-\alpha \end{pmatrix}$$

Los valores propios de esta matriz resultan ser:

$$\{1-\alpha, 1-\alpha(1+\frac{\sqrt{2}}{20}), 1-\alpha(1-\frac{\sqrt{2}}{20})\}$$

por tanto $\rho(D^{-1}(D-\alpha A)) < 1$ si y sólo si el módulo de cada uno de estos valores propios es menor que 1, lo cual ocurre si y sólo si se verifican simultáneamente las tres desigualdades siguientes:

$$\begin{array}{c} \mid 1-\alpha \mid < 1 \Longleftrightarrow 0 < \alpha < 2 \\ \mid 1-\alpha(1+\frac{\sqrt{2}}{20}) \mid < 1 \Longleftrightarrow 0 < \alpha < \frac{2}{1+\frac{\sqrt{2}}{20}} = 1{,}867918234937377 \\ \mid 1-\alpha(1-\frac{\sqrt{2}}{20}) \mid < 1 \Longleftrightarrow 0 < \alpha < \frac{2}{1-\frac{\sqrt{2}}{20}} = 2{,}152182267575185 \end{array}$$

o sea si y sólo si $0 < \alpha < 1{,}867918234937377$.

Obsérvese que para $\alpha = 1$ el método se reduce al método de Jacobi, que es convergente para todo sistema lineal que tenga a esta matriz como matriz de coeficientes.

8. Resolver aproximadamente el sistema lineal $4x+y+z=0, x+4y+z=1, x+y+4z=0$ utilizando el método de Gauss-Seidel comenzando por $(0,0,0)$, estimar el error de la tercera aproximación.

Solución. El método de Gauss-Seidel aplicado a este problema, llamando $\bar{x}^{(k)} = (x^{(k)}, y^{(k)}, z^{(k)})$ a la aproximación k-sima de la solución buscada, equivale a realizar el algoritmo

$$\begin{array}{rcl} x^{(k+1)} & = & -(y^{(k)}+z^{(k)})/4 \\ y^{(k+1)} & = & (1-x^{(k+1)}-z^{(k)})/4 \\ z^{(k+1)} & = & -(x^{(k+1)}+y^{(k+1)})/4 \end{array}$$

comenzando a iterar por $x^{(0)} = y^{(0)} = z^{(0)} = 0$, como nos piden, obtenemos para las tres primeras aproximaciones

$$\bar{x}^{(1)} = \begin{pmatrix} x^{(1)} \\ y^{(1)} \\ z^{(1)} \end{pmatrix} = \begin{pmatrix} 0{,}0 \\ 0{,}25 \\ -0{,}0625 \end{pmatrix}$$

$$\bar{x}^{(2)} = \begin{pmatrix} x^{(2)} \\ y^{(2)} \\ z^{(2)} \end{pmatrix} = \begin{pmatrix} -0{,}046875 \\ 0{,}27734375 \\ -0{,}0576171875 \end{pmatrix}$$

$$\bar{x}^{(3)} = \begin{pmatrix} x^{(3)} \\ y^{(3)} \\ z^{(3)} \end{pmatrix} = \begin{pmatrix} -0{,}054931640625 \\ 0{,}27813720703125 \\ -0{,}055801391601563 \end{pmatrix}$$

que, como es sabido, es equivalente a realizar el algoritmo

$$\bar{x}^{(k+1)} = E_{GS}\bar{x}^{(k)} + F$$

siendo $E_{GS} = (D-L)^{-1}U$ y $F = (D-L)^{-1}b$, donde la matriz de iteración es

$$E_{GS} = \begin{pmatrix} 4 & 0 & 0 \\ 1 & 4 & 0 \\ 1 & 1 & 4 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & -1 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0{,}0 & -0{,}25 & -0{,}25 \\ 0{,}0 & 0{,}0625 & -0{,}1875 \\ 0{,}0 & 0{,}046875 & 0{,}109375 \end{pmatrix}$$

Puesto que la $\| E_{GS} \|_\infty =$ máximo de las normas subuno de sus vectores fila, se tiene que $\| E_{GS} \|_\infty = \text{máx}\{0{,}5, 0{,}25, 0{,}15625\} = 0{,}5 = k < 1$, y dado que $\rho(E_{GS}) < \| E_{GS} \|_\infty = 0{,}5 < 1$ el método es convergente, cosa que ya sabíamos por ser la matriz de coeficientes estrictamente diagonal dominante. Ahora, podemos acotar el error de la tercera aproximación en la forma

$$\| \bar{x}^{(3)} - \bar{x} \|_\infty \leq \frac{0{,}5}{1-0{,}5} \| \bar{x}^{(3)} - \bar{x}^{(2)} \|_\infty = 0{,}008056640625$$

donde $\bar{x}$ representa la solución exacta.

9. Dado el sistema lineal: $9x - 2y = 5, -2x + 4y - z = 1, -y + z = 5/6$, hallar la norma matricial $\| E_R \|_\infty$ de la matriz de iteración del método de relajación con $\omega = 1{,}2$, ¿es convergente dicho método?, acotar el error de la cuarta iteración de dicho método si se comienza a iterar por $(0, 0, 0)$.

Solución. La matriz de iteración del método de relajación, E_R, teniendo en cuenta que $\omega = 1{,}2 = \frac{6}{5}$, está dada por

$$E_R = (\frac{1}{\omega}D - L)^{-1}(\frac{1-\omega}{\omega}D + U) = \begin{pmatrix} \frac{15}{2} & 0 & 0 \\ -2 & \frac{10}{3} & 0 \\ 0 & -1 & \frac{5}{6} \end{pmatrix}^{-1} \cdot \begin{pmatrix} -\frac{3}{2} & 2 & 0 \\ 0 & -\frac{2}{3} & 1 \\ 0 & 0 & -\frac{1}{6} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{2}{15} & 0 & 0 \\ \frac{2}{25} & \frac{3}{10} & 0 \\ \frac{12}{125} & \frac{9}{25} & \frac{6}{5} \end{pmatrix} \cdot \begin{pmatrix} -\frac{3}{2} & 2 & 0 \\ 0 & -\frac{2}{3} & 1 \\ 0 & 0 & -\frac{1}{6} \end{pmatrix} = \begin{pmatrix} -\frac{1}{5} & \frac{4}{15} & 0 \\ -\frac{3}{25} & -\frac{1}{25} & \frac{3}{10} \\ -\frac{18}{125} & -\frac{6}{125} & \frac{4}{25} \end{pmatrix}$$

$$= \begin{pmatrix} -0{,}2 & 0{,}26666666666667 & 0 \\ -0{,}12 & -0{,}04 & 0{,}3 \\ -0{,}144 & -0{,}048 & 0{,}16 \end{pmatrix}$$

por tanto $\rho(E_R) \leq \| E_R \|_\infty = \text{máx}\{0{,}4\bar{6}, 0{,}46, 0{,}352\} = 0{,}4\bar{6} < 1$, luego el método es convergente.

Las aproximaciones sucesivas del método de relajación, que en este problema denotamos por $\bar{x}^{(k)}$, vimos que podían calcularse mediante el algoritmo

$$(\frac{1}{\omega}D - L)\bar{x}^{(k+1)} = (\frac{1-\omega}{\omega}D + U)\bar{x}^{(k)} + b$$

que en componentes adopta la forma general

$$\frac{1}{\omega}a_{ii}\bar{x}_i^{(k+1)} + \sum_{j=1}^{i-1} a_{ij}\bar{x}_j^{(k+1)} = \frac{1-\omega}{\omega}a_{ii}\bar{x}_i^{(k)} - \sum_{j=i+1}^{n} a_{ij}\bar{x}_j^{(k)} + b_i$$

o también

$$\bar{x}_i^{(k+1)} = \frac{\omega}{a_{ii}}(-\sum_{j=1}^{i-1} a_{ij}\bar{x}_j^{(k+1)} + \frac{1-\omega}{\omega}a_{ii}\bar{x}_i^{(k)} - \sum_{j=i+1}^{n} a_{ij}\bar{x}_j^{(k)} + b_i)$$

ahora, en el caso que nos ocupa es $\bar{x} \equiv (x_1, x_2, x_3) \equiv (x, y, z)$ y $\omega = 1{,}2$, con lo que el algoritmo se reduce al siguiente

$$\begin{aligned}
x^{(k+1)} &= \frac{1{,}2}{9}(-\frac{0{,}2}{1{,}2}9x^{(k)} + 2y^{(k)} + 5) \\
y^{(k+1)} &= \frac{1{,}2}{4}(2x^{(k+1)} - \frac{0{,}2}{1{,}2}4y^{(k)} + z^{(k)} + 1) \\
z^{(k+1)} &= 1{,}2(y^{(k+1)} - \frac{0{,}2}{1{,}2}z^{(k)} + \frac{5}{6})
\end{aligned}$$

y partiendo de $x^{(0)} = y^{(0)} = z^{(0)} = 0$, reteniendo tan sólo 6 cifras significativas en los cálculos, se obtienen las cuatro primeras aproximaciones:

$$\bar{x}^{(1)} = \begin{pmatrix} 0{,}666667 \\ 0{,}7 \\ 1{,}84 \end{pmatrix}, \bar{x}^{(2)} = \begin{pmatrix} 0{,}72 \\ 1{,}144 \\ 2{,}0048 \end{pmatrix}$$

$$\bar{x}^{(3)} = \begin{pmatrix} 0{,}827733 \\ 1{,}16928 \\ 2{,}00218 \end{pmatrix}, \bar{x}^{(4)} = \begin{pmatrix} 0{,}812928 \\ 1{,}15455 \\ 1{,}98503 \end{pmatrix}$$

Podemos acotar el error absoluto de la cuarta iteración en la forma

$$\| \bar{x}^{(4)} - \bar{x} \|_\infty \leq \frac{\| E_R \|_\infty}{1 - \| E_R \|_\infty} \| \bar{x}^{(4)} - \bar{x}^{(3)} \|_\infty < 2 \cdot 10^{-2}$$

10. Dados los sistemas lineales: a) $2x + 4y = 8, x - y = -5$, y b) $2x - y = -8, x + y = -1$; probar que el método de Gauss-Seidel es divergente para a) y convergente para b), para este hallar la norma matricial $\| E_{GS} \|_2$ de la matriz de iteración del método de Gauss-Seidel y acotar el error de la sexta iteración de dicho método si se comienza a iterar por $(0, 0)$.

Solución. Como es sabido la matriz de iteración del método de Gauss-Seidel está dada por

$$E_{GS} = (D - L)^{-1}U$$

que para el caso a) resulta ser

$$E_{GS} = \begin{pmatrix} 2 & 0 \\ 1 & -1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & -4 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & -1 \end{pmatrix} \cdot \begin{pmatrix} 0 & -4 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 \\ 0 & -2 \end{pmatrix}$$

cuyos valores propios son $\{-2, 0\}$, por tanto $\rho(E_{GS}) = 2 > 1$ y el método diverge.

Ahora, la matriz de iteración para el caso b) es

$$E_{GS} = \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ -\frac{1}{2} & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & -\frac{1}{2} \end{pmatrix}$$

siendo ahora sus valores propios $\{-\frac{1}{2}, 0\}$, se deduce que $\rho(E_{GS}) = \frac{1}{2} < 1$ y el método converge.

11. Dado el sistema lineal: $7x-2y+3z = 0, -x+5y-2z = 11, x+y-3z = 6$, se pide:

- Averiguar si los métodos de Jacobi y Gauss-Seidel son convergentes.
- Hallar las normas matriciales $\| E_J \|_p$ (con $p = 1, \infty$) de la matriz de iteración del método de Jacobi y acotar el error en la cuarta iteración obtenida por dicho método partiendo del vector inicial nulo.
- Hallar la matriz de iteración del método de Gauss-Seidel (E_{GS}) y su $\| E_{GS} \|_\infty$.
- Por el método de Gauss-Seidel, obtener la cuarta iteración y acotar su error si se comienza a iterar por $(0, 1, 0)$.

Solución. La respuesta a la primera cuestión es afirmativa, ambos métodos convergen en este caso por ser la matriz del sistema estrictamente diagonal dominante ($7 > 2+3 = 5$, $5 > 1+2 = 3$ y $3 > 1+1 = 2$),

lo que constituye una condición suficiente de convergencia para ambos métodos.

Para la segunda cuestión, calculemos en primer lugar la matriz de iteración del método de Jacobi, dada por $E_J = D^{-1}(L+U)$ o sea

$$E_J = \begin{pmatrix} \frac{1}{7} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & -\frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} 0 & 2 & -3 \\ 1 & 0 & 2 \\ -1 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{2}{7} & -\frac{3}{7} \\ \frac{1}{5} & 0 & \frac{2}{5} \\ \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}$$

Por tanto, para las normas pedidas se obtienen

$$\| E_J \|_1 = \text{máx}\{\frac{1}{5}+\frac{1}{3}, \frac{2}{7}+\frac{1}{3}, \frac{3}{7}+\frac{2}{5}\} = \frac{29}{35}$$

$$\| E_J \|_\infty = \text{máx}\{\frac{2}{7}+\frac{3}{7}, \frac{1}{5}+\frac{2}{5}, \frac{1}{3}+\frac{1}{3}\} = \frac{5}{7}$$

ambas son menores que 1. Ahora hemos de obtener la cuarta iteración por el método de Jacobi partiendo del vector inicial nulo, como el algoritmo de este método está dado por las fórmulas

$$\begin{aligned} x^{(k+1)} &= (2y^{(k)} - 3z^{(k)})/7 \\ y^{(k+1)} &= (11 + x^{(k)} + 2z^{(k)})/5 \\ z^{(k+1)} &= (-6 + x^{(k)} + y^{(k)})/3 \end{aligned}$$

comenzando a iterar por $x^{(0)} = y^{(0)} = z^{(0)} = 0$, como nos piden, obtenemos para las cuatro primeras aproximaciones

$$\bar{x}^{(1)} = \begin{pmatrix} x^{(1)} \\ y^{(1)} \\ z^{(1)} \end{pmatrix} = \begin{pmatrix} 0{,}0 \\ 2{,}2 \\ -2 \end{pmatrix}$$

$$\bar{x}^{(2)} = \begin{pmatrix} x^{(2)} \\ y^{(2)} \\ z^{(2)} \end{pmatrix} = \begin{pmatrix} 1{,}4857143 \\ 1{,}4 \\ -1{,}2666667 \end{pmatrix}$$

$$\bar{x}^{(3)} = \begin{pmatrix} x^{(3)} \\ y^{(3)} \\ z^{(3)} \end{pmatrix} = \begin{pmatrix} 0{,}9428572 \\ 1{,}9904762 \\ -1{,}0380952 \end{pmatrix}$$

$$\bar{x}^{(4)} = \begin{pmatrix} x^{(4)} \\ y^{(4)} \\ z^{(4)} \end{pmatrix} = \begin{pmatrix} 1{,}0136054 \\ 1{,}9733334 \\ -1{,}0222222 \end{pmatrix}$$

que, como es sabido, es equivalente a realizar el algoritmo

$$\bar{x}^{(k+1)} = E_J \bar{x}^{(k)} + F$$

siendo $E_J = D^{-1}(L+U)$ y $F = D^{-1}b$. Para acotar el error de la cuarta iteración utilizaremos la $\parallel E_J \parallel_\infty$, resultando

$$\parallel \bar{x}^{(4)} - \bar{x} \parallel_\infty \leq \frac{\frac{5}{7}}{1-\frac{5}{7}} \parallel \bar{x}^{(4)} - \bar{x}^{(3)} \parallel_\infty \leq \frac{5}{2} \mid x^{(4)} - x^{(3)} \mid < 0{,}18$$

En la tercera cuestión nos piden la matriz de iteración del método de Gauss-Seidel, E_{GS}, y su $\parallel E_{GS} \parallel_\infty$, dadas por

$$E_{GS} = \begin{pmatrix} 7 & 0 & 0 \\ -1 & 5 & 0 \\ 1 & 1 & -3 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & 2 & -3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{7} & 0 & 0 \\ \frac{1}{35} & \frac{1}{5} & 0 \\ \frac{2}{35} & \frac{1}{15} & -\frac{1}{3} \end{pmatrix} \cdot \begin{pmatrix} 0 & 2 & -3 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{2}{7} & -\frac{3}{7} \\ 0 & \frac{2}{35} & \frac{11}{35} \\ 0 & \frac{4}{35} & -\frac{4}{105} \end{pmatrix}$$

y

$$\parallel E_{GS} \parallel_\infty = \text{máx}\{\frac{2}{7} + \frac{3}{7}, \frac{2}{35} + \frac{11}{35}, \frac{4}{35} + \frac{4}{105}\} = \frac{5}{7}$$

Para responder a la cuarta cuestión, tengamos en cuenta que el método de Gauss-Seidel aplicado a este problema, llamando $\bar{x}^{(k)} = (x^{(k)}, y^{(k)}, z^{(k)})$ a la aproximación k-sima de la solución buscada, equivale a realizar el algoritmo

$$\begin{aligned} x^{(k+1)} &= (2y^{(k)} - 3z^{(k)})/7 \\ y^{(k+1)} &= (11 + x^{(k+1)} + 2z^{(k)})/5 \\ z^{(k+1)} &= (-6 + x^{(k+1)} + y^{(k+1)})/3 \end{aligned}$$

comenzando a iterar por $x^{(0)} = 0$, $y^{(0)} = 1$ y $z^{(0)} = 0$, como nos piden y realizando los cálculos con MAXIMA, obtenemos para las cuatro primeras aproximaciones:

$$\bar{x}^{(1)} = \begin{pmatrix} x^{(1)} \\ y^{(1)} \\ z^{(1)} \end{pmatrix} = \begin{pmatrix} 0{,}28571428571429 \\ 2{,}257142857142857 \\ -1{,}152380952380952 \end{pmatrix}$$

$$\bar{x}^{(2)} = \begin{pmatrix} x^{(2)} \\ y^{(2)} \\ z^{(2)} \end{pmatrix} = \begin{pmatrix} 1{,}138775510204082 \\ 1{,}966802721088436 \\ -0{,}96480725623583 \end{pmatrix}$$

$$\bar{x}^{(3)} = \begin{pmatrix} x^{(3)} \\ y^{(3)} \\ z^{(3)} \end{pmatrix} = \begin{pmatrix} 0{,}97543245869776 \\ 2{,}009163589245222 \\ -1{,}005134650685671 \end{pmatrix}$$

$$\bar{x}^{(4)} = \begin{pmatrix} x^{(4)} \\ y^{(4)} \\ z^{(4)} \end{pmatrix} = \begin{pmatrix} 1{,}004818732935351 \\ 1{,}998909886312802 \\ -0{,}99875712691728 \end{pmatrix}$$

Finalmente, para acotar el error de la cuarta iteración obtenida, utilizaremos la $\| E_{GS} \|_\infty$, resultando

$$\| \bar{x}^{(4)} - \bar{x} \|_\infty \leq \frac{\frac{5}{7}}{1-\frac{5}{7}} \| \bar{x}^{(4)} - \bar{x}^{(3)} \|_\infty \leq \frac{5}{2} \mid x^{(4)} - x^{(3)} \mid < 0{,}08$$

3.6. Algunos programas Maxima para la resolución de sistemas lineales

3.6.1. Normas vectoriales y matriciales

Para vectores podemos definir las normas usuales por medio de las funciones

```
(%i1) norma_1(a):=apply("+",abs(a));
      norma_2(a):=sqrt(apply("+",abs(a)^2));
      norma_inf(a):=lmax(abs(a));
```

(%o1) $\text{norma_1}(a) := \text{apply}(+, |a|)$

(%o2) $\text{norma_2}(a) := \sqrt{\text{apply}(+, |a|^2)}$

(%o3) $\text{norma_inf}(a) := \text{lmax}(|a|)$

```
(%i4) a:[-5,3,4,-7];norma_1(a);norma_2(a);norma_inf(a);
```

(%o4) $[-5, 3, 4, -7]$

(%o5) 19

(%o6) $3\sqrt{11}$

(%o7) 7

En tanto que para matrices cuadradas A, el radio espectral y las normas matriciales compatibles con las anteriores, podemos obtenerlas con las funciones:

```
(%i8) re(A):=lmax(abs(eigenvalues(A)[1]))$
```

```
(%i9)  norma_1(A):=block(n:matrix_size(A)[1],
       for j:1 thru n do
       c[j]:sum(abs(A[i,j]),i,1,n),
       c:makelist(c[j],j,1,n),
       print("norma_1(A) = ",lmax(c))
       )$
```

```
(%i10) norma_2(A):=print("norma_2(A)
        =",float(sqrt(re(transpose(A).A))))$
```

```
(%i11) norma_inf(A):=block(n:matrix_size(A)[1],
       for i:1 thru n do
       f[i]:sum(abs(A[i,j]),j,1,n),
       f:makelist(f[i],i,1,n),
       print("norma_inf(A) = ",lmax(f))
       )$
```

```
(%i12) A:matrix([2,2,4],[2,3,6],[1,1,1]);print("Radio espectral
        de A = ",float(re(A)))$ norma_1(A)$
        norma_2(A)$ norma_inf(A)$
```

(%o12) $\begin{pmatrix} 2 & 2 & 4 \\ 2 & 3 & 6 \\ 1 & 1 & 1 \end{pmatrix}$

Radio espectral de A = 6,418832675970043

$norma_1(A) = 11$

$norma_2(A) = 8{,}676601657378294$

$norma_\infty(A) = 11$

Observación. Téngase en cuenta que la función "eigenvalues" llama a "solve" y puede que no sepa resolver muchas ecuaciones algebraicas de orden igual o superior a 5, por ello en ese caso habrá que determinar el polinomio característico y resolverlo con otros comandos, por ejemplo "algsys". Lo vemos seguidamente.

```
(%i17) Re(A):=lmax(abs(map(rhs,allroots(charpoly(A,x)))));
```

(%o17) $\mathrm{Re}\,(A) := \mathrm{lmax}\,(|\mathrm{map}\,(rhs, \mathrm{allroots}\,(\mathrm{charpoly}\,(A, x)))|)$

```
(%i18) A:matrix([10,-3,2,1,1],[1,8,-1,3,0],[0,1,12,-5,0],
       [2,-2,3,20,1],[-2,-1,1,3,15]); print("Re(A) = ",Re(A))$
```

$$(\%o18)\quad \begin{pmatrix} 10 & -3 & 2 & 1 & 1 \\ 1 & 8 & -1 & 3 & 0 \\ 0 & 1 & 12 & -5 & 0 \\ 2 & -2 & 3 & 20 & 1 \\ -2 & -1 & 1 & 3 & 15 \end{pmatrix} Re(A) = 17{,}13121755587384$$

3.6.2. Método iterativo de Jacobi

Dentro de los métodos iterativos para la resolución de sistemas de ecuaciones lineales, el más clásico es el método de Jacobi. En las siguientes órdenes vamos a ver cómo aplicar dicho método. Consideramos el sistema $Ax = b$ donde la matriz de coeficientes A y la de términos independientes son

```
(%i20) A:matrix([1,6,-1,2],[4,1,0,1],[0,-1,10,4],[3,-1,-2,7]);
       b:matrix([6],[6],[-2],[19]);
```

$$(\%o20)\quad \begin{pmatrix} 1 & 6 & -1 & 2 \\ 4 & 1 & 0 & 1 \\ 0 & -1 & 10 & 4 \\ 3 & -1 & -2 & 7 \end{pmatrix}$$

$$(\%o21)\quad \begin{pmatrix} 6 \\ 6 \\ -2 \\ 19 \end{pmatrix}$$

Debido a que el método de Jacobi converge para matrices de coeficientes estrictamente diagonal dominantes, vamos a intercambiar la primera con la segunda ecuación (filas), es decir definimos las nuevas matrices A y b

```
(%i22) A:rowswap(A,1,2);
       b:rowswap(b,1,2);
```

$$(\%o22)\quad \begin{pmatrix} 4 & 1 & 0 & 1 \\ 1 & 6 & -1 & 2 \\ 0 & -1 & 10 & 4 \\ 3 & -1 & -2 & 7 \end{pmatrix}$$

$$(\%o23)\quad \begin{pmatrix} 6 \\ 6 \\ -2 \\ 19 \end{pmatrix}$$

Ahora inicializamos las variables, el número máximo de iteraciones "nmax", la dimensión s del sistema y los valores iniciales de las incógnitas, que tomamos iguales a cero

```
(%i24) n_max:25;
       s:matrix_size(b);
       s:s[1];
       P:zerofor(b);
       n:1;
       z:matrix([0],[0],[0],[0]);
```

(%o24) 25
(%o25) $[4, 1]$
(%o26) 4
(%o27) $\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$
(%o28) 1
(%o29) $\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$

Seguidamente realizamos el algoritmo de Jacobi nmax veces partiendo de (0,0,0,0,0)

```
(%i30) while (n<=n_max) do (
             for i:1 thru s do (
             z[i]:1/A[i,i]*(b[i]-sum(A[i,j]*P[j],j,1,i-1)
                                 -sum(A[i,j]*P[j],j,i+1,s))
                                  ),
                    P:z,
                    n:n+1
                         )$
```

Finalmente, mostramos la última aproximación de la solución obtenida

```
(%i31) print("La solución aproximada tras ", nmax, " pasos es ",
        float(z))$
```

La solución aproximada tras 25 pasos es $\begin{pmatrix} 1{,}0 \\ 1{,}6336667969710926 \cdot 10^{-15} \\ -1{,}0 \\ 2{,}0 \end{pmatrix}$

que, prácticamente, coincide con la exacta dada por $x = (1, 0, -1, 2)$.

3.6.3. Método iterativo de Gauss-Seidel

Ahora presentamos un block para resolver un sistema lineal $Ax = b$, por el método de Gauss-Seidel, en el supuesto de que sea convergente al aplicarlo a este sistema, con un número máximo de iteraciones ("nmax") prefijado, que se puede cambiar para obtener una aproximación adecuada.

```
(%i32) /* Método de Gauss-Seidel con nmax iteraciones para
          resolver un sistema de n ecuaciones lineales*/
       GS(A,b,nmax):=block([numer],numer:true,nmax,
       n:matrix_size(A)[1], x:matrix([0],[0],[0],[0],[0],[0]),
       for k:1 thru nmax do (for i:1 thru n do
       (x[i]:(b[i]-sum(A[i,j]*x[j],j,1,i-1)
       -sum(A[i,j]*x[j],j,i+1,n))/A[i,i])),
       print("La solución aproximada tras ",nmax," pasos es ",x))$
```

que ahora aplicamos al caso del sistema lineal $Ax = b$, de seis ecuaciones dado por las matrices A y b que siguen

```
(%i33) A: matrix([10,-3,2,1,1,0],[1,8,-1,3,0,1],[0,1,12,-5,0,2],
       [2,-2,3,20,1,1],[-2,-1,1,3,15,2],[1,-2,1,-1,1,9]);
       b: matrix([16],[-1],[0],[29],[38],[31]);GS(A,b,25)$
```

$$(\%o33) \quad \begin{pmatrix} 10 & -3 & 2 & 1 & 1 & 0 \\ 1 & 8 & -1 & 3 & 0 & 1 \\ 0 & 1 & 12 & -5 & 0 & 2 \\ 2 & -2 & 3 & 20 & 1 & 1 \\ -2 & -1 & 1 & 3 & 15 & 2 \\ 1 & -2 & 1 & -1 & 1 & 9 \end{pmatrix}$$

$$(\%o34) \quad \begin{pmatrix} 16 \\ -1 \\ 0 \\ 29 \\ 38 \\ 31 \end{pmatrix}$$

$$\text{La solución aproximada tras 25 pasos es } \begin{pmatrix} 0{,}99999999999997 \\ -0{,}99999999999997 \\ -1{,}0491607582707729\,10^{-14} \\ 1{,}00000000000001 \\ 2{,}000000000000002 \\ 3{,}000000000000012 \end{pmatrix}$$

que es una buena aproximación a la solución exacta $x = (1, -1, 0, 1, 2, 3)$.

Ejercicio. Mejorar el programa anterior de manera que pare la ejecución si la norma subinfinito del vector diferencia de dos iteraciones consecutivas es menor que un épsilon dado.

3.7. Problemas y trabajos propuestos

Problemas propuestos:

1. Resolver, utilizando los métodos de Gauss y Gauss-Jordan, el sistema lineal

$$\begin{aligned} 2x + 4y + z &= 4 \\ 2x + 6y - z &= 10 \\ x + 5y + 2z &= 2 \end{aligned}$$

2. Dadas las matrices

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \text{ y } B = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Calcular:

a) Las normas $\| \cdot \|_1, \| \cdot \|_2$ y $\| \cdot \|_\infty$ de las matrices A y B.

b) El radio espectral de dichas matrices.

3. Dada la matriz

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 2 & 1 \\ -1 & 0 & 0 \end{pmatrix}$$

hallar sus valores propios, su radio espectral y $\| A \|_2$.

4. Determinar el número de condición de la matriz $A = \begin{pmatrix} 1 & 2 \\ 3 & 7 \end{pmatrix}$ con respecto a las normas matriciales $\| \cdot \|_1, \| \cdot \|_2$ y $\| \cdot \|_\infty$.

5. Aplicando el método de Crout, resolver el sistema

$$\begin{aligned} 2x + 3y - z &= 4 \\ x - y + 3z &= -4 \\ y - z &= 2 \end{aligned}$$

6. Hallar si es posible la factorización de Cholesky de la matriz

$$A = \begin{pmatrix} 13 & 11 & 11 \\ 11 & 13 & 11 \\ 11 & 11 & 13 \end{pmatrix}$$

7. Dado el sistema

$$\begin{aligned} 9x - 2y &= 5 \\ -2x + 4y - z &= 1 \\ -y + z &= -5/6 \end{aligned}$$

Partiendo de $(0, 0, 0)$, obtener un valor aproximado de su solución aplicando seis veces el método de Jacobi, trabajando con cuatro cifras decimales redondeadas. Hacer lo mismo pero utilizando el método de Gauss-Seidel.

8. Aplicar el método de Gauss-Seidel para obtener la solución exacta del sistema

$$\begin{aligned} x + 6y + 2z &= 15 \\ x + y - 6z &= -3 \\ 6x + y + z &= 9 \end{aligned}$$

realizando los cálculos con tres cifras decimales redondeadas (la solución exacta se obtiene tras cinco iteraciones).

9. Probar que el sistema

$$\begin{aligned} x - y &= 1 \\ x - 1{,}01y &= 0 \end{aligned}$$

está mal condicionado, resolviéndolo y comparándolo con la solución del sistema

$$\begin{aligned} x - y &= 1 \\ x - 0{,}99y &= 0 \end{aligned}$$

Hallar el número de condición para el sistema inicial.

10. Hacer dos iteraciones, partiendo de $(1, 1, 1)$ por el método de relajación, tomando $\omega = 1/2$ para aproximar la solución del sistema

$$\begin{aligned} 10x - y - z &= 13 \\ x + 10y + z &= 36 \\ -x - y + 10z &= 35 \end{aligned}$$

Averiguar si el método es convergente en este caso.

11. Sea $Ax = b$ un sistema lineal de orden n, con A real no singular, y sean $x^{(k+1)} = Bx^{(k)} + c$ y $x^{(k+1)} = Cx^{(k)} + d$ dos métodos iterativos consistentes con el sistema dado. Sea entonces $y^{(k)}$ la sucesión de valores dada por el algoritmo $y^{(2k+1)} = By^{(k)} + c$, $y^{(2k+2)} = Cy^{(k)} + d$ (con $y^{(0)}$ arbitrario, $k = 1, 2, \ldots, n$). Probar que si llamamos $x^{(k)} = y^{(2k)}$, $k \geq 0$, las $x^{(k)}$ constituyen una sucesión de vectores que proporciona un nuevo método iterativo $x^{(k+1)} = Gx^{(k)} + e$. Indicar quienes son G y e, y probar que este nuevo método iterativo es consistente con el sistema dado.

12. Dado el sistema lineal

$$\begin{aligned} 4x + 3y &= 24 \\ 3x + 4y - z &= 30 \\ -y + 4z &= -24 \end{aligned}$$

realizar cuatro iteraciones por el método de Jacobi y por el de relajación (con $\omega = 1{,}25$), partiendo en ambos casos de $(1, 1, 1)$. Acotar el error de la cuarta iteración obtenida por el método de Jacobi.

13. Realizar tres iteraciones por los métodos de Jacobi y Gauss-Seidel, comenzando por $(0, 0, 0)$, para aproximar la solución del sistema

$$\begin{aligned} 7x - 2y + 3z &= 0 \\ -x + 5y - 2z &= 11 \\ x + y - 3z &= 6 \end{aligned}$$

Justificar la convergencia de dichos métodos y acotar el error de la tercera aproximación en ambos.

14. Para el sistema lineal $Ax = b$, se forma la sucesión dada por

$$x^{(k+1)} = x^{(k)} - \theta(Ax^{(k)} - b), \text{ con } x^{(0)} \in \mathbb{R}^n \text{ y } \theta \in \mathbb{R}$$

Se pide

- Expresar esta sucesión en la forma $x^{(k+1)} = Ex^{(k)} + F$, dando expresiones explícitas de E y F.
- Dar una condición necesaria y suficiente para que dicho método sea convergente.

Trabajos propuestos:
Se propone en este tema realizar alguno de los siguientes trabajos:

- Métodos directos por bloques.
- Métodos iterativos para sistemas tridiagonales.
- Métodos iterativos por bloques.
- Método del gradiente conjugado.
- Métodos tipo gradiente para la resolución de sistemas no lineales.

Capítulo 4

Interpolación polinomial. Aproximación por mínimos cuadrados

4.1. Interpolación

4.1.1. Introducción: diferentes problemas de interpolación

El problema de la interpolación consiste en calcular el valor de una función en un punto, cuando o bien no se conoce la expresión explícita de la misma, o bien no es fácil de evaluar dicha función en ese punto. Para resolverlo, se construye una función fácil de evaluar y que coincida con la función objeto del problema en los datos que conocemos sobre esta.

En todo problema de interpolación hay que concretar dos cuestiones básicas: 1º) los datos que se desea que sean comunes a la función dada y a la que la va a interpolar, y 2º) el tipo de función que se va a utilizar como función de interpolación.

Los tipos más usuales de problemas de interpolación son los que se describen a continuación.

1. **Interpolación polinomial de Lagrange.** El problema de la interpolación polinomial de Lagrange consiste en dados los valores de una función f, $f_i = f(x_i)$ $(i = 0, 1, 2, \dots, n)$, en $n + 1$ puntos distintos $\{x_i\}_0^n$ del intervalo $[a, b]$, determinar, si existe, un polinomio $p(x)$ de grado menor o igual a n tal que $p(x_i) = f_i$ $(i = 0, 1, 2, \dots, n)$, a dicho polinomio se le llama el **polinomio de interpolación** de f en los $n+1$ puntos dados.

2. **Interpolación de Taylor.** En este problema se suponen conocidos los valores de la función f y sus derivadas sucesivas hasta el orden n en el punto x_0 y se trata de hallar un polinomio $p(x)$ de grado menor o igual que n tal que $p^{(k)}(x_0) = f^{(k)}(x_0)$ $(k = 0, 1, 2, \ldots, n)$.

3. **Interpolación de Hermite.** Ahora se suponen conocidos los valores de la función f y su derivada f' en los puntos $x_0, x_1, \ldots, x_n$, que abreviamos por f_i y f'_i (para $i = 0, 1, 2, \ldots, n$) y se trata de hallar un polinomio $p(x)$, de grado menor o igual a $2n + 1$ tal que $p(x_i) = f_i$ y $p'(x_i) = f'_i$ (para $i = 0, 1, 2, \ldots, n$).

4. **Interpolación trigonométrica.** En este caso se conocen los valores de f en los $2n+1$ puntos distintos $x_0, x_1, \ldots, x_{2n}$ del intervalo $[-\pi, \pi)$ y se trata de hallar un polinomio trigonométrico de grado n de la forma

$$p_n(x) = a_0 + \sum_{k=1}^{n}(a_k coskx + b_k senkx)$$

tal que $p_n(x_i) = f_i$ (para $i = 0, 1, 2, \ldots, 2n$).

Aunque la existencia y unicidad de solución de cada uno de los problemas planteados se puede resolver de un modo similar, nos vamos a referir solamente al primero de ellos en los párrafos que siguen.

4.1.2. Interpolación polinomial de Lagrange

En primer lugar veamos el siguiente teorema que garantiza la existencia y unicidad de solución del problema de interpolación de Lagrange planteado anteriormente.

Teorema 26 *Existe un único polinomio de grado a lo más n tal que* $p(x_i) = f_i$ *para* $i = 0, 1, \ldots, n$.

Demostración. Veamos en primer lugar la **unicidad** de solución. Suponer que $p(x)$ y $q(x)$ son dos polinomios de grado a lo más n, verificando dichas condiciones, entonces el polinomio diferencia $r(x) = p(x) - q(x)$ es también un polinomio de grado a lo más n y además verifica que $r(x_i) = p(x_i) - q(x_i) = f_i - f_i = 0$ para $i = 0, 1, \ldots, n$, luego $r(x)$ posee $n + 1$ raíces distintas lo cual implica que $r(x) = 0$, ya que por el teorema fundamental del álgebra todo polinomio no nulo de grado n tiene exactamente n raíces, contando cada una tantas veces como indique su multiplicidad, luego posee a lo más n raíces distintas, en consecuencia debe de ser $r(x) = 0$ y por tanto $p(x) = q(x)$ como queríamos demostrar.

Probaremos la **existencia** de modo constructivo, para ello consideremos los **polinomios de base de Lagrange** de grado n, $l_k(x)$ para $k = 0, 1, \ldots, n$ definidos por

$$\boxed{l_k(x) = \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{(x - x_i)}{(x_k - x_i)} = \frac{(x - x_0)\cdots(x - x_{k-1})(x - x_{k+1})\cdots(x - x_n)}{(x_k - x_0)\cdots(x_k - x_{k-1})(x_k - x_{k+1})\cdots(x_k - x_n)}}$$

entonces, es fácil ver que $l_k(x_i) = 0$ para todo $i \neq k$ y $l_k(x_k) = 1$. Luego el polinomio de interpolación buscado, $p(x)$, se obtiene como combinación lineal de los polinomios de base de Lagrange $l_k(x)$ en la forma

$$\boxed{p(x) = f_0 l_0(x) + f_1 l_1(x) + \cdots + f_n l_n(x) = \sum_{k=0}^{n} f_k l_k(x)}$$

que se conoce como **fórmula de Lagrange** para dicho polinomio.

Para el caso $n = 1$ se tiene la **interpolación lineal**, que consiste en hacer pasar una poligonal por los puntos dados. Por otro lado, la fórmula anterior se simplifica si los puntos de interpolación, también llamados nodos están igualmente separados.

El principal inconveniente de la fórmula de Lagrange estriba en que al añadir nuevos puntos de interpolación no nos sirven los cálculos anteriores, siendo necesario rehacer todos los cálculos comenzando por los nuevos polinomios de base de Lagrange, que serán de un grado superior; por ello, será conveniente disponer de otra fórmula para determinar el polinomio de interpolación, que permita una más fácil transición del polinomio $p_{k-1}(x)$ que interpola a f en $x_0, x_1, \ldots, x_{k-1}$ al $p_k(x)$ que lo hace en los puntos $x_0, x_1, \ldots, x_{k-1}, x_k$, lo que se aborda en el párrafo siguiente.

4.1.3. Diferencias divididas: fórmula de Newton

Con objeto de obtener una fórmula que cumpla el objetivo marcado al final del párrafo anterior, veamos lo que ocurre al pasar del polinomio p_{k-1} que interpola a f en los nodos $x_0, x_1, \ldots, x_{k-1}$ al p_k que la interpola en los puntos $x_0, x_1, \ldots, x_{k-1}, x_k$, el polinomio diferencia $q_k(x) = p_k(x) - p_{k-1}(x)$ es un polinomio de grado menor o igual que k, que se anula para los puntos $x_0, x_1, \ldots, x_{k-1}$, ya que en dichos puntos se tiene $p_k(x_i) = p_{k-1}(x_i) = f(x_i)$ para $i = 0, 1, \ldots, k-1$, luego debe ser de la forma

$$q_k(x) = A_k(x - x_0)(x - x_1)\cdots(x - x_{k-1}) = A_k \prod_{i=0}^{k-1}(x - x_i)$$

con A_k constante. Por tanto resulta que

$$p_k(x) = p_{k-1}(x) + q_k(x) = p_{k-1}(x) + A_k \prod_{i=0}^{k-1}(x - x_i)$$

de donde se deduce que

$$A_k = \frac{p_k(x) - p_{k-1}(x)}{\prod_{i=0}^{k-1}(x - x_i)}$$

y dándole a x el valor x_k, como $p_k(x_k) = f_k$ resulta para $k \geq 1$ que

$$\boxed{A_k = \frac{f_k - p_{k-1}(x_k)}{\prod_{i=0}^{k-1}(x_k - x_i)}}$$

dado que $p_0(x) = f(x_0)$, podemos definir $A_0 = f(x_0)$ y con la fórmula anterior se pueden calcular sin dificultad los coeficientes $A_0, A_1, \ldots, A_n$; al coeficiente A_k se le llama **diferencia dividida de f de orden k en los puntos** $x_0, x_1, \ldots, x_k$ y se representa por

$$\boxed{A_k = f[x_0, x_1, \ldots, x_k]}$$

o también en la forma

$$\boxed{A_k = [x_0, x_1, \ldots, x_k]f}$$

con ayuda de estos coeficientes y utilizando la primera notación para las diferencias divididas, el polinomio de interpolación de f en los nodos $x_0, x_1, \ldots, x_n$ puede escribirse en la forma

$$p_n(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)(x-x_1) + \cdots + A_n(x-x_0)(x-x_1)\cdots(x-x_{n-1})$$

o también

$$\boxed{p_n(x) = \sum_{i=0}^{n} f[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1}(x - x_j)}$$

que se denomina **fórmula de Newton en diferencias divididas**. Como se ve en la fórmula para pasar de p_{k-1} a p_k es suficiente con añadir un término a la primera. Por otro lado, para obtener dicha fórmula se requiere el cálculo previo de las diferencias divididas de todos los órdenes de f en los puntos $x_0, x_1, \ldots, x_n$; lo que se ve facilitado por medio de las dos proposiciones siguientes cuya demostración no haremos, pero que puede verse en el libro de M. Gasca: Cálculo Numérico, editado por la UNED.

Proposición 3 *Para* $k \geq 1$ *se verifica que*

$$f[x_0, x_1, \ldots, x_k] = \sum_{i=0}^{k} \frac{f(x_i)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_k)}$$

por tanto es una función simétrica de sus argumentos, es decir que

$$f[x_0, x_1, \ldots, x_k] = f[x_{j_0}, x_{j_1}, \ldots, x_{j_k}]$$

para toda permutación $(j_0, j_1, \ldots, j_k)$ *de los índices* $(0, 1, \ldots, k)$.

Proposición 4 *Para* $k \geq 1$ *se verifica que*

$$f[x_0, x_1, \ldots, x_{k-1}, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}$$

4.1.4. Diferencias finitas: fórmula de Newton

Definición 9 *Sea una función* f *definida sobre una sucesión de puntos igualmente espaciados* $x_j = x_0 + jh$, $j \in \mathbb{Z}$ *y* $h > 0$, *se llama* ***diferencia progresiva de*** f ***en*** x_k *a*

$$\triangle f(x_k) = f(x_k + h) - f(x_k) = f(x_{k+1}) - f(x_k)$$

para simplificar denotemos $f_j = f(x_j)$, *con lo cual se tiene*

$$\triangle f_k = f_{k+1} - f_k$$

Las diferencias progresivas de orden superior se definen por inducción en la forma

$$\triangle^{n+1} f_k = \triangle(\triangle^n f_k) = \triangle^n f_{k+1} - \triangle^n f_k, \ \forall n \geq 1$$

además se conviene que $\triangle^0 f_k = f_k$, *a* $\triangle^n f_k$ *se le llama la* ***diferencia progresiva de orden n de*** f ***en*** x_k.

Análogamente se definen las diferencias regresivas.

Definición 10 *En las condiciones anteriores, se llama* ***diferencia regresiva de*** f ***en*** x_k *a*

$$\nabla f_k = f_k - f_{k-1}$$

y se definen las diferencias regresivas de orden superior en la forma

$$\nabla^{n+1} f_k = \nabla(\nabla^n f_k) = \nabla^n f_k - \nabla^n f_{k-1}, \ \forall n \geq 1$$

además, se conviene que $\nabla^0 f_k = f_k$, *a* $\nabla^n f_k$ *se le llama la* ***diferencia regresiva de orden n de*** f ***en*** x_k.

La relación entre ambas diferencias está dada por

$$\nabla f_k = \triangle f_{k-1}$$

luego

$$\nabla^n f_k = \triangle^n f_{k-n}$$

La siguiente proposición nos proporciona la relación entre las diferencias progresivas y las diferencias divididas en el caso de puntos igualmente espaciados.

Proposición 5 *Para todo* $n \geq 0$ *se verifica*

$$\boxed{\triangle^n f_k = n! h^n f[x_k, x_{k+1}, \ldots, x_{k+n}]}$$

Demostración. Se hará por inducción, para $n = 0$ es evidente pues

$$\triangle^0 f_k = f_k = f(x_k) = f[x_k]$$

supongamos que dicha propiedad es cierta para $n = 0, 1, \ldots, r$. Veamos que también lo es para $n = r + 1$, en efecto: $\triangle^{r+1} f_k = \triangle(\triangle^r f_k) = \triangle^r f_{k+1} - \triangle^r f_k = r!h^r f[x_{k+1}, x_{k+2}, \ldots, x_{k+r+1}] - r!h^r f[x_k, x_{k+1}, \ldots, x_{k+r}] = r!h^r(x_{k+r+1} - x_k) f[x_k, x_{k+1} \ldots, x_{k+r}, x_{k+r+1}]$, luego la propiedad es cierta para todo $n \in \mathbb{N}$.

Análogamente, para diferencias regresivas se tiene

$$\boxed{\forall n \geq 0 : \nabla^n f_k = n! h^n f[x_{k-n}, x_{k-n+1}, \ldots, x_{k-1}, x_k]}$$

Fórmulas de Newton progresiva y regresiva

Según vimos, la fórmula de Newton para el polinomio de interpolación de f en $n+1$ puntos distintos $x_0, x_1, \ldots, x_n$, se podía escribir en la forma

$$p(x) = \sum_{i=0}^{n} f[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j)$$

Ahora bien, si los puntos están igualmente espaciados, o sea $x_j = x_0 + jh$, en virtud de la proposición 3 anterior se tiene

$$f[x_0, \ldots, x_i] = \frac{\triangle^i f_0}{i! h^i}$$

con lo que el polinomio de interpolación puede escribirse en la forma

$$p(x) = \sum_{i=0}^{n} \frac{\triangle^i f_0}{i! h^i} \prod_{j=0}^{i-1} (x - x_j)$$

y haciendo el cambio de variable $x = x_0 + th$ nos queda

$$p(x) = p(x_0 + th) = \sum_{i=0}^{n} \frac{\triangle^i f_0}{i!} \prod_{j=0}^{i-1}(t - j)$$

por lo que si definimos el número combinatorio generalizado

$$\binom{t}{i} = \frac{t(t-1)\cdots(t-i+1)}{i!}$$

nos queda

$$\boxed{p(x) = p(x_0 + th) = \sum_{i=0}^{n} \binom{t}{i} \triangle^i f_0}$$

conocida como **fórmula de Newton progresiva**.

Análogamente, si escribimos el polinomio $p(x)$ que interpola a f en los puntos $x_n, x_{n-1}, \ldots, x_0$ en la forma de Newton

$$p(x) = \sum_{i=0}^{n} f[x_n, x_{n-1}, \ldots, x_{n-i}] \prod_{j=0}^{i-1}(x - x_{n-j})$$

y para puntos igualmente separados, haciendo el cambio $x = x_n + th$, podemos escribir el polinomio de interpolación en la forma

$$\boxed{p(x) = p(x_n + th) = \sum_{i=0}^{n} \binom{t+i-1}{i} \nabla^i f_n = \sum_{i=0}^{n} (-1)^i \binom{-t}{i} \nabla^i f_n}$$

que se conoce como **fórmula de Newton regresiva**.

4.1.5. Estimación del error de interpolación

Dada una función real f, definida en el intervalo $[a, b]$, sea $p_n(x)$ el polinomio de interpolación de f en los puntos distintos $x_0, x_1, \ldots, x_n$ del intervalo $[a, b]$ (que si no hay lugar a confusión denotamos simplemente como $p(x)$); entonces, dado cualquier otro punto $x \in [a, b]$ se define el **error de interpolación** como $E(x) = f(x) - p(x)$. En los teoremas siguientes se dan sendas expresiones del error que nos permitirán en muchos casos su acotación.

Teorema 27 *Si f es una función definida en los puntos $x_0, x_1, \ldots, x_n, x$ todos distintos, y $p_n(x)$ es el polinomio de grado menor o igual que n que*

interpola a f en los puntos $x_0, x_1, \ldots, x_n$, el error de interpolación se puede escribir en la forma

$$\boxed{E(x) = f(x) - p_n(x) = f[x_0, x_1, \ldots, x_n, x] \prod_{i=0}^{n} (x - x_i)}$$

Demostración. Como hemos visto para la diferencia dividida se tiene

$$f[x_0, x_1, \ldots, x_k] = \frac{f_k - p_{k-1}(x_k)}{\prod_{i=0}^{k-1} (x_k - x_i)} \ (k \geq 1)$$

Entonces, tomando $k = n + 1$ y llamando $x_{n+1} = x$ se tendrá

$$f(x) - p_n(x) = f[x_0, x_1, \ldots, x_n, x] \prod_{i=0}^{n} (x - x_i)$$

como queríamos demostrar.

La fórmula del teorema anterior para el error de interpolación no deja de ser una expresión puramente formal, con objeto de dar otra, que sirva para el cometido de acotar errores, es necesario dar condiciones de regularidad sobre la función f, lo que se contempla en el teorema siguiente.

Teorema 28 *Si $f \in \mathbb{C}^{(n)}([a, b])$ y $x_0, x_1, \ldots, x_n$ son $n + 1$ puntos distintos de $[a, b]$, entonces existe un punto $\xi \in (a, b)$ tal que*

$$\boxed{f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}}$$

(De hecho es $a \leq min(x_0, x_1, \ldots, x_n) < \xi < max(x_0, x_1, \ldots, x_n) \leq b$).

Demostración. La función error de interpolación de f en los $n+1$ puntos dados, $E(x) = f(x) - p_n(x)$, es de clase $\mathbb{C}^{(n)}([a, b])$ por ser diferencia de dos funciones que lo son, y se anula en los puntos $x_0, x_1, \ldots, x_n$; entonces, aplicando el teorema de Rolle, resulta que $E'(x)$ se anulará en, al menos, n puntos distintos intermedios entre aquellos, siendo $E'(x) \in \mathbb{C}^{(n-1)}([a, b])$, aplicando nuevamente el teorema de Rolle $E''(x)$ se anulará en, al menos, $n - 1$ puntos distintos intermedios entre los anteriores, así se llega a que $E^{(n-1)}(x)$ es de clase $\mathbb{C}^{(1)}([a, b])$ y se anula en, al menos, dos puntos en (a, b), luego $E^{(n)}(x)$ se anula en, al menos, un punto $\xi \in (a, b)$, es decir $E^{(n)}(\xi) = f^{(n)}(\xi) - p_n^{(n)}(\xi) = 0$, ahora bien la derivada enésima de un polinomio es el coeficiente de su término de mayor grado multiplicado por $n!$, por tanto $p_n(x) = f[x_0.x_1, \ldots, x_n]n!$, pues el coeficiente de x^n en la fórmula de Newton del polinomio de interpolación es la diferencia dividida $f[x_0.x_1, \ldots, x_n]$, de todo lo cual se deduce que $f[x_0.x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}$.

Corolario 1 *Si $f \in \mathbb{C}^{(n+1)}([a,b])$, $x_0, x_1, \ldots, x_n, x$ son $n+2$ puntos distintos en $[a,b]$ y $p_n(x)$ es el polinomio de interpolación de f en $x_0, x_1, \ldots, x_n$, entonces*

$$\boxed{E(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i)}$$

con ξ intermedio entre $x_0, x_1, \ldots, x_n, x$.

Expresión muy útil para acotar el error de interpolación como se pone de manifiesto en el corolario siguiente.

Corolario 2 *Si llamamos $M = \max_{a \leq t \leq b} \mid f^{(n+1)}(t) \mid$ se tendrá la siguiente acotación:*

$$\boxed{\mid E(x) \mid \leq \frac{M}{(n+1)!} \max_{a \leq t \leq b} \mid \prod_{i=0}^{n} (t - x_i) \mid \ , \ \forall x \in [a,b]}$$

4.1.6. Funciones splines. Splines cúbicos

Las funciones **splines** están definidas a trozos por varios polinomios, de manera que se unen entre sí mediante ciertas condiciones de continuidad, este tipo de funciones suele dar mejor resultado a la hora de aproximar la forma de una función dada que tomar polinomios de grado alto, que suele producir grandes oscilaciones. Veamos a continuación la definición precisa de spline.

Definición 11 *Sea $a = t_0 < t_1 < \ldots < t_n = b$ una partición del intervalo $[a,b]$ se llama **spline de grado** $m \in \mathbb{N}$ con nudos en $t_0 < t_1 < \ldots < t_n$ a una función $s(x)$ definida en $[a,b]$ verificando*

1. *La restricción de $s(x)$ a cada $[t_i, t_{i+1}]$, $i = 0, 1, \ldots, n-1$ es un polinomio $s_i(x)$ de grado menor o igual que m.*

2. *$s(x)$ y sus derivadas de órdenes sucesivas hasta el orden $m-1$ inclusive son continuas en $[a,b]$.*

Ejemplos de funciones splines son las poligonales, definidas por funciones lineales en cada subintervalo y continuas en $[a,b]$, que constituyen splines de grado 1. Las funciones splines de grado impar tienen buenas propiedades de suavidad en sus gráficas, ahora bien, dado que las poligonales ya son discontinuas en su derivada primera, estamos interesados en la interpolación por splines cúbicos, muy utilizados en la práctica.

Splines cúbicos

Para interpolar f sobre los puntos $a = t_0 < t_1 < \ldots < t_n = b$ mediante un **spline cúbico**, sean $f_i = f(t_i)$, $h_i = t_{i+1} - t_i$ y $s_i(x)$ la restricción de $s(x)$ al subintervalo $[t_i, t_{i+1}]$. Como $s_i(x)$ es un polinomio de grado tres, su derivada segunda s_i'' es un polinomio de primer grado, o sea se trata de una recta; entonces, denotando por z_i, z_{i+1} a los valores, desconocidos de momento, de $s_i''(t_i)$ y $s_i''(t_{i+1})$ respectivamente, la ecuación de dicha recta (que pasa por dos puntos (t_i, z_i) y (t_{i+1}, z_{i+1})) podrá escribirse en la forma:

$$s_i''(x) = z_{i+1}\frac{x - t_i}{h_i} + z_i\frac{t_{i+1} - x}{h_i}$$

e integrando dos veces se obtiene

$$s_i(x) = \frac{z_{i+1}}{6h_i}(x - t_i)^3 + \frac{z_i}{6h_i}(t_{i+1} - x)^3 + E_i x + F_i$$

Ahora, el término de primer orden con coeficientes constantes $E_i x + F_i$, puede escribirse en la forma $E_i x + F_i = C_i(x - t_i) + D_i(t_{i+1} - x)$ y de las condiciones $s_i(t_i) = f_i$, $s_i(t_{i+1}) = f_{i+1}$, se deducen los valores de C_i y D_i, quedando el spline en la forma

$$s_i(x) = \frac{z_{i+1}}{6h_i}(x-t_i)^3+\frac{z_i}{6h_i}(t_{i+1}-x)^3+(\frac{f_{i+1}}{h_i}-\frac{z_{i+1}h_i}{6})(x-t_i)+(\frac{f_i}{h_i}-\frac{z_i h_i}{6})(t_{i+1}-x)$$

expresión que garantiza, por un lado, la continuidad del spline s y, por otro, que este coincide con f en los nudos $t_0 < t_1 < \ldots < t_n$; para que s' sea continua en $[t_0, t_n]$ debe verificarse que $s_{i-1}'(t_i) = s_i'(t_i)$ para $i = 1, \ldots, n-1$. Entonces, derivando la expresión del spline $s_i(x)$, así como la de $s_{i-1}(x)$ y evaluando para $x = t_i$, se tiene que las z_i han de verificar el sistema

$$h_{i-1}z_{i-1} + 2(h_{i-1} + h_i)z_i + h_i z_{i+1} = \frac{6}{h_i}(f_{i+1} - f_i) - \frac{6}{h_{i-1}}(f_i - f_{i-1}) \ (*)$$

para $i = 1, 2, \ldots, n-1$, así pues los valores $z_0, z_1, \ldots, z_n$ deben verificar las $n-1$ ecuaciones anteriores. Por otro lado, la continuidad de $s''(x)$ está asegurada pues se tiene, por construcción, que $s_{i-1}''(t_i) = z_i = s_i''(t_i)$. Luego si se fijan arbitrariamente z_0 y z_n en (*) resulta un sistema de $n-1$ ecuaciones lineales con $n-1$ incógnitas, que denotando por

$$b_i = \frac{6}{h_i}(f_{i+1} - f_i) - \frac{6}{h_{i-1}}(f_i - f_{i-1})$$

se puede escribir como el sistema lineal, con matriz de coeficientes tridiagonal, simétrica y estrictamente diagonal dominante, que sigue

$$\begin{aligned} 2(h_0 + h_1)z_1 + h_1 z_2 &= b_1 - h_0 z_0 \\ h_1 z_1 + 2(h_1 + h_2)z_2 + h_2 z_3 &= b_2 \\ &\vdots \\ h_{n-2}z_{n-2} + 2(h_{n-2} + h_{n-1})z_{n-1} &= b_{n-1} - h_{n-1}z_n \end{aligned}$$

si hacemos $z_0 = z_n = 0$ se obtiene el denominado **Spline natural**.

4.2. Introducción al problema de la mejor aproximación

Con objeto de hablar de la mejor aproximación de un elemento f de un espacio V, mediante un elemento f_s de un subconjunto S de V, es necesario dotarnos de un útil matemático para "medir" la distancia entre dos elementos, de hecho nos serviremos de la noción de espacio vectorial normado.

Definición 12 *Se llama espacio normado a un espacio vectorial real o complejo V dotado de una norma $\| \cdot \|$.*

Seguidamente citamos algunos de los espacios normados usuales:

1. $\mathbb{R}$ con la norma $\| x \| = |x|$ (norma del valor absoluto).

2. $\mathbb{R}^n$ y $\mathbb{C}^n$ con las normas $\| x \|_p = (\sum_{i=1}^n \mid x_i \mid^p)^{\frac{1}{p}}$ para $1 \leq p < \infty$ y $\| x \|_\infty = \max\{\mid x_i \mid \mid 1 \leq i \leq n\}$ (denominada norma del supremo); casos especialmente interesantes son los relativos a $p = 1$ y $p = 2$ dados por $\| x \|_1 = (\sum_{i=1}^n \mid x_i \mid)$ y $\| x \|_2 = (\sum_{i=1}^n \mid x_i \mid^2)^{\frac{1}{2}} = (\bar{x}^t x)^{1/2}$, aquí x denota la matriz columna con las mismas componentes que el vector x, en tanto que $\bar{x}^t$ es la matriz traspuesta de la conjugada de x, es decir la matriz fila cuyas componentes son las conjugadas de las de x (esta es denominada norma euclídea del vector x).

3. $\mathcal{C}([a, b])$ (espacio vectorial de las funciones reales continuas definidas en el intervalo $[a, b]$) con la norma $\| f \| = max_{a \leq x \leq b} |f(x)|$, denominada norma de la convergencia uniforme.

4. $\mathcal{C}([a, b])$ con la norma $\| f \| = (\int_a^b |f(x)|^2 dx)^{\frac{1}{2}}$, denominada norma de la convergencia cuadrática.

Definición 13 *Dado en espacio vectorial normado V, $x_1, x_2, \cdots, x_m$, m vectores linealmente independientes de V e $y \in V$ (un vector cualquiera), el problema fundamental de la* ***aproximación lineal*** *consiste en aproximar y por una combinación lineal $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m$ con α_i escalares tales que $\| y - (\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m) \|$ sea lo menor posible. Se denomina* ***mejor aproximación*** *de y por una combinación lineal de $x_1, x_2, \cdots, x_m$, a un elemento $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m$ tal que*

$$\| y - (\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m) \| \leq \| y - (\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m) \|$$

para todo $\beta_i \in \mathbb{K}$ (aquí $\mathbb{K}$ es el cuerpo de los números reales o complejos).

Ejercicio. El problema de la mejor aproximación depende del conjunto que utilicemos para aproximar y del útil que utilicemos para medir la distancia. A modo de ejemplo, el alumno puede tratar de buscar la mejor aproximación de la función definida en el intervalo $[0, 1]$ por $f(x) = x^4$, mediante polinomios de grado 1, $p(x) = ax + b$, de manera que

1. $\int_0^1 (x^4 - p(x))^2 dx$ sea mínimo.
2. $|p(0)| + |1 - p(1)|$ sea mínimo.
3. $max_{0 \leq x \leq 1} |x^4 - p(x)|$ sea mínimo.

Se nos plantean tres problemas en relación con la mejor aproximación lineal en un espacio normado: 1) ¿Cómo caracterizarla?, 2) Averiguar si existe y es única, y finalmente, 3) ¿Cómo calcularla en el caso de que exista?. A la segunda pregunta responde en parte el siguiente teorema (cuya demostración es sencilla pero la omitimos para abreviar).

Teorema 29 *En un espacio normado V, el problema de la aproximación lineal planteado en la definición anterior posee solución, lo que no podemos asegurar, en general, es la unicidad de la misma.*

4.2.1. Espacios prehilbertianos

Definición 14 *Sea V un espacio vectorial sobre $\mathbb{K}$ (siendo $\mathbb{K}$ el cuerpo de los números reales o complejos) se llama* ***producto escalar o interior*** *sobre V a toda aplicación de $V \times V$ en $\mathbb{K}$ que a cada par (x, y) asocia un escalar, que denotamos por $(x|y) =< x|y >= x \circ y$, verificando las tres propiedades siguientes:*

1. *Es lineal en la primera componente, es decir $\forall \alpha, \beta \in \mathbb{K}$ y $\forall x, y, z \in V$ es $(\alpha x + \beta y|z) = \alpha(x|z) + \beta(y|z)$.*

2. *$\forall x, y \in V$ es $(x|y) = \overline{(y|x)}$, es decir un producto es igual al conjugado del producto que resulta de invertir los factores, es la denominada simetría hermítica, que para el caso real es simplemente la propiedad de simetría.*

3. *$\forall x \in V, x \neq 0$ es $(x|x) > 0$, es decir la aplicación es definida positiva.*

*El par $(V, (|))$ formado por un espacio vectorial real o complejo y un producto escalar se denomina espacio **pre-Hilbert o prehilbertiano**. Si el cuerpo $\mathbb{K} = \mathbb{R}$ se dice **espacio euclídeo**, y si es $\mathbb{K} = \mathbb{C}$, se dice **espacio unitario o hermítico**.*

Ejemplos. Veamos tres ejemplos de espacios prehilbertianos.

1. $\mathbb{R}^n$ con el producto escalar definido $\forall x, y \in \mathbb{R}^n$ por $(x|y) = \sum_{i=1}^{n} x_i y_i$ es un espacio euclídeo de dimensión n.

2. $\mathbb{C}^n$ con el producto escalar definido $\forall x, y \in \mathbb{C}^n$ por $(x|y) = \sum_{i=1}^{n} x_i \overline{y_i}$ es un espacio unitario de dimensión n.

3. $\mathcal{C}([a, b], \mathbb{K})$ es el espacio vectorial de las funciones continuas de $[a, b]$ en $\mathbb{K}$, definiendo el producto escalar en la forma $(f|g) = \int_a^b f(x)\overline{g(x)}dx$ se convierte en un espacio prehilbertiano (euclídeo si el cuerpo es real y unitario si es complejo).

Definición 15 *Dado un espacio prehilbertiano $(V, (|))$, $\forall x \in V$ se define su norma en la forma $\parallel x \parallel = \sqrt{(x|x)}$ (es la norma asociada al producto escalar). Es fácil ver que es una norma. Además, con el producto escalar del ejemplo 1) anterior, $\forall x \in \mathbb{R}^n$ es $\parallel x \parallel = \sqrt{\sum_{i=1}^{n} x_i^2}$; en tanto que con el producto escalar del ejemplo 2), $\forall x \in \mathbb{C}^n$ es $\parallel x \parallel = \sqrt{\sum_{i=1}^{n} |x_i|^2}$.*

4.2.2. Ortogonalidad. Bases ortonormales

Definición 16 *Dos vectores x e y de un espacio prehilbertiano V se dicen **ortogonales** si $(x|y) = 0$. Un sistema de vectores de V se dice **ortogonal** si sus vectores son ortogonales dos a dos, si además cada vector del sistema es unitario (es decir de norma 1) el sistema se dice ortonormal. Es inmediato probar que todo sistema ortogonal de vectores no nulos es libre.*

Recordamos ahora el teorema de Gram-Schmidt que es de utilidad para construir sistemas ortonormales a partir de un sistema libre.

Teorema 30 *Sea $\{u_n\}$ una sucesión de elementos de V linealmente independientes, entonces la sucesión $\{v_n\}$ dada por el algoritmo*

$$\boxed{\begin{array}{c} v_1 = u_1 \\ v_{n+1} = u_{n+1} - \sum_{i=1}^{n} \frac{(u_{n+1}|v_i)}{\|v_i\|^2} v_i \end{array}}$$

es una sucesión de vectores ortogonales y la sucesión $\{w_n\} = \{\frac{v_n}{\|v_n\|}\}$ *es un sistema ortonormal; además*

$$\mathbb{K}(u_1, \cdots, u_n) = \mathbb{K}(v_1, \cdots, v_n) = \mathbb{K}(w_1, \cdots, w_n)$$

Ahora, resulta inmediato probar la siguiente proposición.

Proposición 6 *Sea V prehilbertiano de dimensión finita n; entonces la coordenada i-ésima de un vector $v \in V$ respecto de una base ortonormada $\{e_1, e_2, \cdots, e_n\}$ es igual al producto escalar de v por e_i, es decir $\alpha_i = (v|e_i)$.*

Definición 17 *Sea S un subespacio vectorial de un espacio prehilbertiano V,* ***un vector*** *$x \in V$* ***es ortogonal a*** *S y se indica $v \perp S$ si $\forall y \in S$ es $(x|y) = 0$. Dado un subespacio vectorial S de V se llama* ***suplemento ortogonal*** *de S y se indica por $S^\perp$ al conjunto $S^\perp = \{x \in V | x \perp S\}$. Es inmediato ver que $S^\perp$ es un subespacio vectorial de V y que si S está finitamente engendrado, entonces x es ortogonal a S si y sólo si lo es a un sistema generador de S.*

4.2.3. Mejor aproximación por mínimos cuadrados

Teorema 31 ***(Teorema de la proyección)*** *Sea S un subespacio de dimensión finita m de un espacio prehilbertiano V, entonces*

1. *Todo vector $v \in V$ se expresa de modo único en la forma $v = v_s + (v - v_s)$ donde $v_s \in S$ y $v - v_s \in S^\perp$. Es decir $V = S \oplus S^\perp$ (es la suma directa de S y su subespacio ortogonal). A v_s se le llama la proyección ortogonal de v sobre S, y si $\{w_1, w_2, \cdots, w_n\}$ es una base ortonormada de S, entonces se tiene*

$$\boxed{v_s = \sum_{i=1}^{m} (v|w_i) w_i}$$

2. *Además, para cualquier vector $y \in S$, $y \neq v_s$ se verifica que*

$$\| v - v_s \| < \| v - y \|$$

Es decir v_s es el elemento de S que está más próximo a v en la norma asociada al producto escalar, por ello se denomina la ***mejor aproximación por mínimos cuadrados*** *de v mediante elementos de S.*

Observación.-La demostración del teorema anterior es sencilla pero la omitiremos, sólo resaltaremos sus aplicaciones. Por lo reflejado en el mismo, si se quiere hallar la mejor aproximación por mínimos cuadrados de un elemento v de un espacio prehilbertiano V, por elementos de un subespacio S de dimensión finita, basta con hallar primero una base ortonormada de S, utilizando el teorema de Gram-Schmidt, y posteriormente hacer la proyección de v sobre esa base en la forma indicada en el apartado 1 del teorema de la proyección, pues esa es la mejor aproximación por mínimos cuadrados buscada según se refleja en el apartado 2. Dicha mejor aproximación se dice continua si el producto escalar está definido por una integral (ejemplo 3 anterior) y discreta si lo está por una suma (ejemplos 1 y 2).

4.2.4. Mejor aproximación por mínimos cuadrados continua o discreta

Sea $V = \mathcal{C}([a,b])$ el espacio vectorial real de las funciones continuas en el intervalo cerrado y acotado $[a,b]$, si para cada par de funciones f y g de V, definimos el producto escalar en la forma $(f|g) = \int_a^b f(x)g(x)dx$, tendremos un espacio prehilbertiano real. Entonces, dado cualquier subespacio S de V, de dimensión finita m, para toda $f \in V$ existe un único $f_s \in S$ tal que $f - f_s \in S^{\perp}$, es decir existe un único $f_s \in S$ tal que $(f - f_s|\nu) = 0$ para todo $\nu \in S$ y, teniendo en cuenta la definición del producto escalar, podemos decir que para toda $f \in V$ existe un único $f_s \in S$ tal que $\int_a^b (f(x) - f_s(x))\nu(x)dx = 0$ para toda $\nu(x) \in S$, siendo esta f_s **la mejor aproximación por mínimos cuadrados continua** de f por elementos de S. Para hallar f_s caben dos posibilidades:

1. Si $\{\nu_1, \nu_2, \cdots, \nu_m\}$ es una base de S, $f_s = \sum_{i=1}^m \lambda_i \nu_i$ y se calculan los $\lambda_1, \lambda_2, \cdots, \lambda_m$ con la condición de que $f - f_s$ sea ortogonal a S o lo que es lo mismo a $\{\nu_1, \nu_2, \cdots, \nu_m\}$, es decir se han de verificar para $j \in \{1, 2, \cdots, m\}$ las condiciones siguientes:

$$(f - f_s|\nu_j) = 0 \Longleftrightarrow (f|\nu_j) = (f_s|\nu_j)$$

y sustituyendo f_s por su expresión anterior, se obtiene el sistema de m ecuaciones lineales

$$\boxed{\sum_{i=1}^{m} \lambda_i (\nu_i|\nu_j) = (f|\nu_j) \ (j = 1, 2, \cdots, m)}$$

que es compatible determinado (pues su matriz de coeficientes es simétrica y definida positiva), lo que permite obtener las incógnitas $\lambda_1, \cdots, \lambda_m$.

2. Si $\{w_1, w_2, \cdots, w_m\}$ es una base ortonormada de S, entonces f_s viene dada por la fómula
$$\boxed{f_s = \sum_{i=1}^{m} (f|w_j)w_j}$$
y se denomina **suma de Fourier de f** con respecto a la base ortonormada dada.

En cambio, si consideramos en $\mathbb{R}^n$ el producto escalar euclídeo, definido $\forall x, y \in \mathbb{R}^n$ por $(x|y) = \sum_{i=1}^{n} x_i y_i$, dado un subespacio vectorial S de dimensión m, de acuerdo con el teorema de la proyección, para todo $y \in \mathbb{R}^n$ existe un único $y_s \in S$ que es **la mejor aproximación por mínimos cuadrados discreta** de y por elementos de S, que se calcula por cualquiera de las dos formas anteriores, teniendo en cuenta que ahora el producto escalar viene dado por una suma.

4.3. Problemas resueltos

1. Hallar el polinomio de interpolación de Lagrange que pasa por los puntos: $(0, -2), (1, 6), (3, 40)$.

   Solución. Recordemos que por tres puntos distintos dados pasa un único polinomio de grado menor o igual que dos, que puede escribirse en la forma de Lagrange como
$$p(x) = f_0 l_0(x) + f_1 l_1(x) + f_2 l_2(x) = -2l_0(x) + 6l_1(x) + 40l_2(x)$$
   estando dados los polinomios de base de Lagrange por
$$l_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-1)(x-3)}{(0-1)(0-3)} = \frac{x^2-4x+3}{3}$$
$$l_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-0)(x-3)}{(1-0)(1-3)} = \frac{x^2-3x+0}{-2}$$
$$l_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-0)(x-1)}{(3-0)(3-1)} = \frac{x^2-x+0}{6}$$
   con lo cual el polinomio de interpolación de Lagrange resulta ser
$$p(x) = -\frac{2}{3}(x^2-4x+3) - 3(x^2-3x) + \frac{20}{3}(x^2-x) = 3x^2+5x-2$$

2. Los valores de f dados en la tabla siguiente son los de un cierto polinomio de grado cuatro

x_i	0	1	2	3	4
$f(x_i)$	1	5	31	121	341

se pide hallar $f(5)$, ¿cuál es su error?.

Solución. Puesto que nos dan los valores de f en cinco puntos distintos hay un único polinomio $p(x)$, de grado menor o igual que cuatro que pasa por ellos, es su polinomio de interpolación, y al ser $f(x)$ un cierto polinomio de grado cuatro y cumplir esas condiciones coincide con este, en virtud de la unicidad de dicho polinomio, por tanto $f(x) = p(x)$ para todo x, luego $f(5) = p(5)$ y el error sería nulo $E(5) = f(5) - p(5) = 0$. Para calcular el polinomio de interpolación podemos usar distintos métodos, lo haremos en este caso por el método de Newton en diferencias divididas que puede escribirse en la forma

$$\boxed{p(x) = p_4(x) = \sum_{i=0}^{4} f[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1}(x - x_j)}$$

Para ello, calculemos los coeficientes del polinomio de Newton por medio de la tabla de diferencias divididas:

$f[x_0] = \boxed{1}$

$f[x_1] = 5$ $\quad f[x_0, x_1] = \boxed{4}$

$f[x_2] = 31$ $\quad f[x_1, x_2] = 26$ $\quad f[x_0, x_1, x_2] = \boxed{11}$

$f[x_3] = 121$ $\quad f[x_2, x_3] = 90$ $\quad f[x_1, x_2, x_3] = 32$ $\quad f[x_0, x_1, x_2, x_3] = \boxed{7}$

$f[x_4] = 341$ $\quad f[x_3, x_4] = 220$ $\quad f[x_2, x_3, x_4] = 65$ $\quad f[x_1, x_2, x_3, x_4] = 11$

y finalmente se tiene

$$f[x_0, x_1, x_2, x_3, x_4] = \frac{f[x_1, x_2, x_3, x_4] - f[x_0, x_1, x_2, x_3]}{x_4 - x_0} = \frac{11 - 7}{4 - 0} = \boxed{1}$$

donde hemos encuadrado los coeficientes que intervienen en el polinomio de interpolación de Newton en diferencias divididas, que resulta ser

$$p(x) = 1 + 4x + 11x(x-1) + 7x(x-1)(x-2)+$$
$$+x(x-1)(x-2)(x-3) = 1 + x + x^2 + x^3 + x^4$$

Y para el valor pedido se obtiene $f(5) = p(5) = 781$.

3. Obtener, mediante la fórmula de Newton progresiva, el polinomio que interpola a la función de la que se conocen los siguientes datos:

$$\begin{array}{rccccc} x_i: & 0 & 1 & 2 & 3 & 4 \\ f(x_i): & 0 & 1 & 8 & 27 & 64 \end{array}$$

Solución. Puesto que los nodos de interpolación están igualmente espaciados podemos utilizar la fórmula de Newton en diferencias progresivas, dada por la expresión

$$\boxed{p(x) = p(x_0 + th) = \sum_{i=0}^{4} \binom{t}{i} \triangle^i f_0}$$

por ser en este caso $x_0 = 0$ y $h = 1$, resulta $\boxed{t = x}$. Nos queda hallar los coeficientes $\triangle^i f_0$ y teniendo en cuenta que $\triangle^0 f_j = f_j$ lo podemos hacer mediante la siguiente tabla de diferencias progresivas:

$$\begin{array}{lllll} f(x_0) = \boxed{0} & & & & \\ f(x_1) = 1 & \triangle f_0 = \boxed{1} & & & \\ f(x_2) = 8 & \triangle f_1 = 7 & \triangle^2 f_0 = \boxed{6} & & \\ f(x_3) = 27 & \triangle f_2 = 19 & \triangle^2 f_1 = 12 & \triangle^3 f_0 = \boxed{6} & \\ f(x_4) = 64 & \triangle f_3 = 37 & \triangle^2 f_2 = 18 & \triangle^3 f_1 = 6 & \triangle^4 f_0 = \boxed{0} \end{array}$$

donde hemos encuadrado los coeficientes que intervienen en la fórmula de Newton progresiva, que resulta ser

$$p(x) = 0 + \binom{x}{1} + 6\binom{x}{2} + 6\binom{x}{3} =$$
$$= x + 6\frac{x(x-1)}{2} + 6\frac{x(x-1)(x-2)}{6} = \boxed{x^3}$$

4. Calcular, mediante la fórmula de Newton regresiva, el polinomio interpolador que pasa por los puntos: $(-1, 1), (0, 1), (1, 1)$ y $(2, -5)$.

Solución. Para puntos de interpolación igualmente separados, el polinomio $p(x)$ que interpola a f en los puntos x_3, x_2, x_1, x_0 puede escribirse en la forma de Newton regresiva como

$$\boxed{p(x) = p(x_3 + th) = \sum_{i=0}^{3} (-1)^i \binom{-t}{i} \nabla^i f_3}$$

como $x_3 = 2$ y $h = 1$, resulta $t = \frac{x-x_3}{1} = x - 2$, y para obtener dicho polinomio hemos de calcular previamente los coeficientes $\nabla^i f_3$, ahora teniendo en cuenta que $\nabla^0 f_j = f_j$ lo podemos hacer mediante la tabla de diferencias regresivas:

$$
\begin{array}{llll}
f(x_3) = \boxed{-5} & & & \\
f(x_2) = 1 & \nabla f_3 = \boxed{-6} & & \\
f(x_1) = 1 & \nabla f_2 = 0 & \nabla^2 f_3 = \boxed{-6} & \\
f(x_0) = 1 & \nabla f_1 = 0 & \nabla^2 f_2 = 0 & \nabla^3 f_3 = \boxed{-6}
\end{array}
$$

donde hemos encuadrado los coeficientes que intervienen en la fórmula de Newton regresiva, que resulta ser

$$p(x) = -5 + 6\binom{2-x}{1} - 6\binom{2-x}{2} + 6\binom{2-x}{3} =$$

$$= -5+6(2-x)-6\frac{(2-x)(1-x)}{2}+6\frac{(2-x)(1-x)(-x)}{6} = \boxed{1+x-x^3}$$

5. Calcular una aproximación de $\sqrt[3]{2}$ y dar una cota del valor absoluto del error cometido utilizando el polinomio de interpolación de la función $f(x) = 2^x$ en los puntos $\{-1, 0, 1, 2\}$.

   Solución. La función a interpolar es $f(x) = 2^x$ en $\{-1, 0, 1, 2\}$, utilizaremos el polinomio de interpolación de Newton en diferencias divididas, dado ahora por la fórmula

$$p(x) = p_3(x) = \sum_{i=0}^{3} f[x_0, x_1, \dots, x_i] \prod_{j=0}^{i-1} (x - x_j)$$

   Para ello, calculemos los coeficientes del polinomio de Newton por medio de la tabla de diferencias divididas:

$$
\begin{array}{llll}
f[x_0] = \boxed{\frac{1}{2}} & & & \\
f[x_1] = 1 & f[x_0, x_1] = \boxed{\frac{1}{2}} & & \\
f[x_2] = 2 & f[x_1, x_2] = 1 & f[x_0, x_1, x_2] = \boxed{\frac{1}{4}} & \\
f[x_3] = 4 & f[x_2, x_3] = 2 & f[x_1, x_2, x_3] = \frac{1}{2} & f[x_0, x_1, x_2, x_3] = \boxed{\frac{1}{12}}
\end{array}
$$

donde hemos encuadrado los coeficientes que intervienen en el polinomio de interpolación de Newton en diferencias divididas, que resulta ser

$$p_3(x) = \frac{1}{2} + \frac{1}{2}(x+1) + \frac{1}{4}(x+1)x + \frac{1}{12}(x+1)x(x-1)$$

de donde obtenemos la aproximación

$$\sqrt[3]{2} = 2^{\frac{1}{3}} \cong p_3(\frac{1}{3}) = \frac{1}{2} + \frac{2}{3} + \frac{1}{9} - \frac{2}{81} = \frac{203}{162} = \boxed{1{,}253086\cdots}$$

Para acotar el error de esta aproximación, tengamos en cuenta la fórmula del error de interpolación dada en este caso por la expresión

$$E(x) = f(x) - p(x) = \frac{f^{(iv)}(\xi)}{4!}(x+1)x(x-1)(x-2)$$

siendo ξ un punto intermedio entre $\{-1, 0, 1, 2, x\}$. En particular, para el punto $\xi = \frac{1}{3}$ se tendrá

$$\mid E(\frac{1}{3}) \mid = \mid f(\frac{1}{3}) - p(\frac{1}{3}) \mid = \mid \frac{f^{(iv)}(\xi)}{4!} \mid\mid (\frac{1}{3}+1)\frac{1}{3}(\frac{1}{3}-1)(\frac{1}{3}-2) \mid$$

con $\xi \in [-1, 2]$. Puesto que $f(x) = 2^x$, se sigue que $f^{(iv)}(x) = 2^x(\log 2)^4$ (donde log significa logaritmo neperiano) es una función creciente en el intervalo $[-1, 2]$, luego $f^{(iv)}(\xi) < f^{(iv)}(2) = 2^2(\log 2)^4 = 0{,}9233\cdots < 0{,}93$, por tanto

$$\mid E(\frac{1}{3}) \mid < \frac{0{,}93}{24} \cdot \frac{4}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{5}{3} < \boxed{0{,}02}$$

6. Hallar una base ortonormada del espacio euclídeo $\mathbb{R}^3$ a partir de la siguiente $\{(1, 0, -1), (1, 1, 0), (1, 1, 1)\}$ y calcular las coordenads del vector $v = (3, 1, 2)$ en dicha base. Asimismo, sea $\mathbb{R}_2[x]$ el espacio vectorial de los polinomios de grado menor o igual que dos, con coeficientes reales, dotado del producto escalar $(p(x)|q(x)) = \int_0^1 p(x)q(x)dx$, que resulta ser un espacio euclídeo, se desea hallar una base ortonormada partiendo de la base canónica $\{1, x, x^2\}$.

   Solución. (Damos una indicación pero se deben completar los cálculos). Basta aplicar el algoritmo de Gram-Schmidt a la base dada, para obtener una base ortogonal y luego dividiendo cada vector por su módulo, tendremos la base ortonormada $\{w_1, w_2, w_3\}$, que viene dada por los vectores $w_1 = (\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}})$, $w_2 = (\frac{\sqrt{2}}{2\sqrt{3}}, \frac{\sqrt{2}}{\sqrt{3}}, \frac{\sqrt{2}}{2\sqrt{3}})$, $w_3 = (\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$. Ahora, las coordenadas α_i de un vector v en una base ortonormada

viene dadas por $\alpha_i = (v|w_i)$ $(i = 1, 2, 3)$, y en este caso resultan ser $\alpha_1 = \frac{1}{\sqrt{2}}$,$\alpha_2 = \frac{7\sqrt{2}}{2\sqrt{3}}$ y $\alpha_3 = \frac{4\sqrt{3}}{3}$.

Para la segunda parte, basta con aplicar nuevamente el algoritmo de Gram-Schmidt para obtener la siguiente base ortonormada $\{1, \sqrt{3}(2x-1), \sqrt{5}(6x^2 - 6x + 1)\}$.

7. En el espacio euclídeo $\mathcal{C}([-1, 1])$, dotado del producto escalar $(f|g) = \int_0^1 f(x)g(x)dx$, se desea hallar la mejor aproximación por mínimos cuadrados continua de $f(x) = x^{\frac{1}{3}}$ por polinomios de segundo grado.

   Solución. Lo haremos de dos formas diferentes, como se ha visto en teoría.

   a) En primer lugar, una base de los polinomios de grado menor o igual que dos es la canónica $\{\nu_1 = 1, \nu_2 = x, \nu_3 = x^2\}$, entonces si la mejor aproximación buscada es $f_s = \lambda_1\nu_1 + \lambda_2\nu_2 + \lambda_3\nu_3$, los coeficientes λ_i serán la solución del sistema lineal de mínimos cuadrados dado por las ecuaciones:

   $$\boxed{\sum_{i=1}^{3} \lambda_i(\nu_i|\nu_j) = (x^{\frac{1}{3}}|\nu_j)\ (j = 1, 2, 3)}$$

   Una vez hechos los cálculos resultan $\lambda_1 = 0$, $\lambda_2 = \frac{9}{7}$ y $\lambda_3 = 0$, por tanto obtenemos la mejor aproximación $f_s(x) = \frac{9}{7}x$.

   b) Para la segunda forma, hemos de obtener una base ortonormada partiendo de la base canónica, en la forma indicada en el problema anterior, pero con el producto escalar que ahora corresponde, esta resulta ser $\{w_1 = \frac{1}{\sqrt{2}}, w_2 = \frac{\sqrt{3}}{\sqrt{2}}x, w_3 = \frac{\sqrt{5}}{2\sqrt{2}}(3x^2 - 1)$, finalmente

   $$\boxed{f_s = \sum_{i=1}^{3}(x^{\frac{1}{3}}|w_j)w_j = \frac{9}{7}x}$$

   con el mismo resultado anterior.

4.4. Algunos programas Maxima para interpolación y aproximación de funciones

4.4.1. Cálculo directo de polinomios de interpolación

Como es sabido, dados los valores de una función $f(x)$, $fi = f(x_i)$ en $n + 1$ puntos distintos $\{x_i|i = 0, 1, 2, \cdots, n\}$ del intervalo $[a, b]$, existe un

único polinomio $p(x)$ de grado menor o igual a n tal que $p(x_i) = fi$ para $i = 0, 1, 2, \cdots, n$, a dicho polinomio se le llama el polinomio de interpolación de f en los $n+1$ puntos dados. Puede escribirse como $P(x) = a_0 + a_1 x + ... + a_{n-1}x^{n-1} + a_n x^n$ y hallar los coeficientes $a_0, a_1, \cdots, a_n$ del mismo de manera que se cumplan las $n + 1$ ecuaciones lineales: $P(x_i) = fi$ $(i = 0, 1, \cdots, n)$ siendo el sistema correspondiente compatible determinado, pues el determinante de la matriz de coeficientes es un determinante de Vandermonde, que no se anula por ser los puntos distintos, aunque sabemos que dicha matriz está mal condicionada, de ahí el interés en otros métodos para obtener el polinomio de interpolación para n grande. Pero veamos algún ejemplo con este método. Se desea obtener el polinomio de interpolación que pasa por los puntos: $(0, 1), (1, 5), (2, 31), (3, 121), (4, 341)$, como se trata de 5 puntos, por ellos pasa un único polinomio de grado, a lo más, cuatro de la forma $P(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$, para ello hemos de resolver el sistema lineal:

```
(%i1)  p(x):=a_0+a_1*x+a_2*x^2+a_3*x^3+a_4*x^4;
       linsolve([p(0)=1,p(1)=5,p(2)=31,
       p(3)=121,p(4)=341],[a0,a1,a2,a3,a4]);
```

(%o1) $\mathrm{p}(x) := a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$

(%o2) $[a_0 = 1, a_1 = 1, a_2 = 1, a_3 = 1, a_4 = 1]$

Luego se trata del polinomio $x^4 + x^3 + x^2 + x + 1$.

4.4.2. Fórmula de Newton en diferencias divididas

La fórmula de Newton, en diferencias divididas, del polinomio de interpolacón está dada por la expresión

$$p(x) = \sum_{i=0}^{n} f[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j)$$

siendo las $f[x_0, x_1, ..., x_i]$ las diferencias divididas de f de orden i en los puntos $x_0, x_1, \cdots, x_i$, que se calculan recurrentemente como sigue

$$f[x_0, x_1, ..., x_i] = (f[x_1, ..., x_i] - f[x_0, x_1, ..., x_{i-1}])/(x_i - x_0)$$

con $f[x_k] = f(x_k) = f_k = fk$ y las escribiremos como $f01..i$, para evitar problemas con otras instrucciones de Maxima, asimismo debido al editor de este programa en muchas ocasiones ponemos xi en vez de x_i.

```
(%i3) kill(all)$ x0:0$ x1:1$ x2:2$ x3:3$ x4:4$
      f0:1$ f1:5$ f2:31$ f3:121$ f4:341$
      /*Cálculo de las diferencias divididas*\
      f01:(f1-f0)/(x1-x0)$
      f12:(f2-f1)/(x2-x1)$
      f23:(f3-f2)/(x3-x2)$
      f34:(f4-f3)/(x4-x3)$
      f012:(f12-f01)/(x2-x0)$
      f123:(f23-f12)/(x3-x1)$
      f234:(f34-f23)/(x4-x2)$
      f0123:(f123-f012)/(x3-x0)$
      f1234:(f234-f123)/(x4-x1)$
      f01234:(f1234-f0123)/(x4-x0)$

      p(x):=f0+f01*(x-x0)+f012*(x-x0)*(x-x1)+
      f0123*(x-x0)*(x-x1)*(x-x2)+
      f01234*(x-x0)*(x-x1)*(x-x2)*(x-x3);

      expand(p(x));
```

(%o21)

$$p(x) := f0 + f01(x - x0) + f012(x - x0)(x - x1) +$$
$$+ f0123(x - x0)(x - x1)(x - x2) +$$
$$+ f01234(x - x0)(x - x1)(x - x2)(x - x3)$$

(%o22) $x^4 + x^3 + x^2 + x + 1$

Y vuelve a dar el polinomio hallado antes.

4.4.3. Mejor aproximación por mínimos cuadrados continua

En el espacio euclídeo de las funciones reales continuas definidas en el intervalo $[-\pi, \pi]$, dotado del producto escalar $(f|g) = \int_{-\pi}^{\pi} f(x)g(x)dx$, se pide hallar la mejor aproximación por mínimos cuadrados continua de la función $f(x) = 1 + |x|$ por polinomios trigonométricos de grado menor o igual que dos. Nos piden aproximarla por una función de la forma: $f_s(x) := a_0 + a_1 cos(x) + b_1 sin(x) + a_2 cos(2x) + b_2 sin(2x)$.

Para ello hemos de calcular a_0, a_1, b_1, a_2, b_2 tales que sean soluciones del sistema lineal de mínimos cuadrados (cuyas ecuaciones representan que la

proyección de f sobre el subespacio de los polinomios trigonométricos de grado menor o igual que dos es igual a la de f_s). En el programa que sigue se carga el paquete "abs_integrate" que permite realizar las integrales definidas con módulos de funciones.

```
(%i23) load(abs_integrate)$ numer:true$
       fs(x):=a0+a1*cos(x)+b1*sin(x)+a2*cos(2*x)+b2*sin(2*x);
       f(x):=1+abs(x);
       eq1:integrate(fs(x),x,-%pi,%pi)=
                       integrate(f(x),x,-%pi,%pi);
       eq2:integrate(fs(x)*cos(x),x,-%pi,%pi)=
                       integrate(f(x)*cos(x),x,-%pi,%pi);
       eq3:integrate(fs(x)*sin(x),x,-%pi,%pi)=
                       integrate(f(x)*sin(x),x,-%pi,%pi);
       eq4:integrate(fs(x)*cos(2*x),x,-%pi,%pi)=
                       integrate(f(x)*cos(2*x),x,-%pi,%pi);
       eq5:integrate(fs(x)*sin(2*x),x,-%pi,%pi)=
                       integrate(f(x)*sin(2*x),x,-%pi,%pi);
       linsolve([eq1,eq2,eq3,eq4,eq5],[a0,a1,b1,a2,b2]);
```

(%o25) $\mathrm{fs}\,(x) := a0 + a1\cos(x) + b1\sin(x) + a2\cos(2\,x) + b2\sin(2\,x)$
(%o32) $[a0 = 2{,}570796326794899, a1 = -1{,}273239544735164, b1 = 0, a2 = -7{,}7661728742053211\,10^{-16}, b2 = 0]$

Luego la mejor aproximación pedida será

$$f_s(x) = 2{,}5707963267949 - 1{,}273239544735165cos(x) + \\ -7{,}766172874205326 \cdot 10^{-16}cos(2x)$$

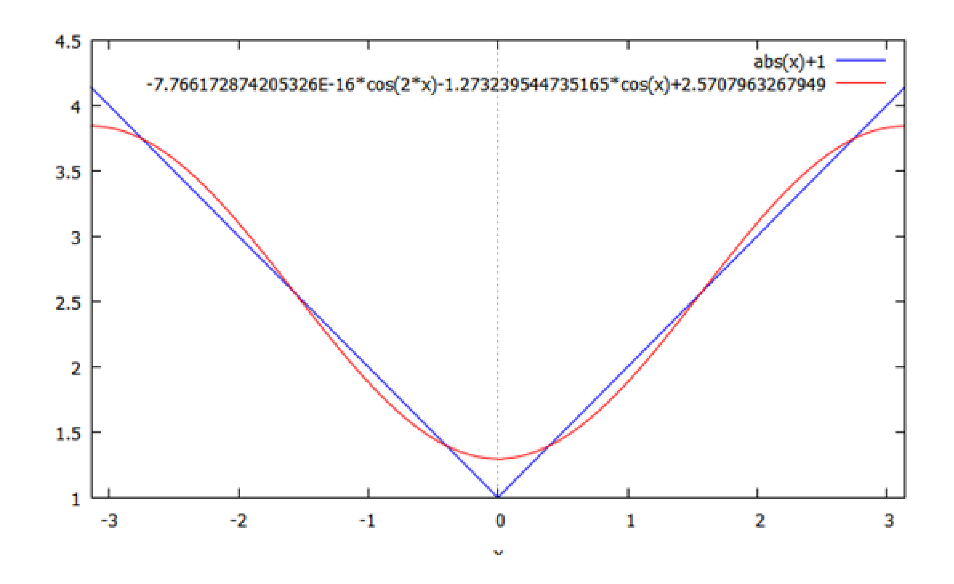
Ahora, representemos ambas funciones en el intervalo $[-\pi, \pi]$

```
(%i33) kill(all,f,fs)$ f(x):=1+abs(x);
       fs(x):=2.5707963267949-1.273239544735165*cos(x)
                 -7.766172874205326*(10^-16)*cos(2*x);
       wxplot2d([f(x),fs(x)],[x,-%pi,%pi]);
```

(%o1) $\mathrm{f}\,(x) := 1 + |x|$
(%o2)

$$f_s(x) := 2{,}5707963267949 - 1{,}273239544735165cos(x) + \\ -7{,}766172874205326 \cdot 10^{-16}cos(2x)$$

(%o3)



4.5. Problemas y trabajos propuestos

Problemas propuestos:

1. Utilizando el polinomio de interpolación de Lagrange, estimar el valor de $f(4)$, sabiendo que $f(-1) = 2$, $f(0) = 0$, $f(3) = 4$ y $f(7) = 7$.

2. Mediante la fórmula de Newton en diferencias divididas, obtener el polinomio de interpolación que se ajusta a la tabla

$x:$	0	1	2
$f(x):$	1	2	3

3. Dada la tabla

$x:$	0,125	0,250	0,375	0,500
$f(x):$	0,792	0,773	0,744	0,704

   - Obtener el polinomio de interpolación mediante la fórmula de Newton progresiva.
   - Obtener el polinomio de interpolación mediante la fórmula de Newton regresiva.
   - Razonar la posible igualdad de ambos.

4. De la tabla de logaritmos decimales se obtienen los siguientes valores aproximados para $log_{10}(x)$: $log_{10}(1{,}5) = 0{,}17609$, $log_{10}(2) = 0{,}30103$, $log_{10}(3) = 0{,}47712$, $log_{10}(3{,}5) = 0{,}54407$. Mediante el polinomio de interpolación aproximar el $log_{10}(2{,}5)$ y acotar el error cometido.

5. Utilizando la fórmula de Newton progresiva, hallar el polinomio de grado menor o igual que 5, que pasa por los puntos de la tabla

$x:$	-3	-1	1	3	5	7
$f(x):$	14	4	2	8	22	44

6. Dada la tabla de valores

$x:$	0	1	2	3	4
$f(x):$	0	1	4	7	8

   Se pide el spline cúbico que pasa por esos puntos y una estimación de $f(1{,}5)$ y de $f'(2)$.

7. Usar una fórmula de Newton progresiva para hallar el polinomio de interpolación que aproxime a una función $f(x)$ tal que $f(0) = 0$ y $f(n) = 1^2 + 2^2 + \cdots + n^2,\ \forall n \in \mathbb{N}$.

8. Probar que si $f(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$, entonces para todo x es $f[x_0, x_1, \ldots, x_n, x] = 1$ (Ayuda: Usar la relación existente entre la diferencia dividida y una adecuada derivada de f).

9. El valor de $e^{0,2}$ se estima por interpolación a partir de los valores $e^0 = 1, e^{0,1} \simeq 1{,}1052$ y $e^{0,3} \simeq 1{,}3499$. Hallar dicho valor interpolado y acotar el error cometido.

10. Se considera el espacio $\mathcal{C}([-1, 1])$, dotado del producto escalar $(f|g) = \int_{-1}^{1} f(x)g(x)dx$, se desea encontrar la mejor aproximación por mínimos cuadrados continua de $f(x) = e^x$, por polinomios de grado menor o igual que 2.

11. En el espacio euclídeo de las funciones reales continuas definidas en el intervalo $[-\pi, \pi]$, dotado del producto escalar $(f|g) = \int_{-\pi}^{\pi} f(x)g(x)dx$, se pide hallar la mejor aproximación por mínimos cuadrados continua de la función $f(x) = 1 + |x|$ por polinomios trigonométricos de grado menor o igual que dos.

12. Hallar la mejor aproximación de la función constante $f(x) = 1$, mediante funciones de la forma $ax + bx^2$, por mínimos cuadrados continua en $[-1, 1]$, y por mínimos cuadrados discreta en $\{-1, 0, 1\}$, definiendo el producto escalar en este caso en la forma $(f(x)|g(x)) = f(-1)g(-1) + f(0)g(0) + f(1)g(1)$.

Trabajos propuestos:
Se propone en este tema realizar alguno de los siguientes trabajos:

- Diferencias divididas con argumentos repetidos y errores en los métodos interpolatorios.
- Interpolación de Hermite y aplicaciones.
- Interpolación por polinomios trigonométricos y transformada rápida de Fourier.
- Mejor aproximación uniforme de una función continua por polinomios de grado n.

Capítulo 5

Derivación e integración numérica

5.1. Derivación numérica

En muchas ocasiones es necesario disponer del valor de la derivada de una función f en un punto c, o bien de su integral en un intervalo $[a, b]$, pero no nos es posible calcularlas por disponer tan sólo de una tabla de valores de la función en cuestión o bien por ser su expresión analítica inmanejable. Entonces, lo más usual es aproximar una u otra por una combinación lineal de los valores de f en los puntos $x_i,\ i = 0, 1, \ldots, n$, en los que f está definida, es decir

$$\boxed{f'(c) \simeq \sum_{i=0}^{n} a_i f(x_i)} \tag{5.1}$$

al tomar este valor como aproximación de la derivada se comete un error de derivación numérica, que denotamos por $EDN(f)$, que está definido por medio de la fórmula

$$\boxed{f'(c) = \sum_{i=0}^{n} a_i f(x_i) + EDN(f)}$$

Nuestro objetivo en este punto será el estudio de fórmulas de este tipo y de su error correspondiente. Diremos que la fórmula (5.1) es **exacta para la función** ϕ si $EDN(\phi) = 0$, y se dirá **exacta de orden p** si $EDN(x^i) = 0$ para $i = 0, 1, \ldots, p$ y $EDN(x^{p+1}) \neq 0$ (si se verifican las primeras diremos que es de orden al menos p). Asimismo, se dirá que es de **tipo interpolatorio** si se ha obtenido derivando el polinomio de interpolación de f; es decir si f es derivable en un punto $c \in [a, b]$ y es $p(x) = \sum_{i=0}^{n} f_i l_i(x)$ el polinomio

de interpolación de f, en los puntos x_i, $(i = 0, 1, \ldots, n)$ de dicho intervalo, expresado en la forma de Lagrange, entonces se tendrá

$$\boxed{f'(c) \simeq p'(c) = \sum_{i=0}^{n} f_i l_i'(c)} \tag{5.2}$$

Puesto que el error de interpolación se define por la igualdad

$$f(x) = p(x) + E(x)$$

si f es derivable en c, $E(x)$ también lo será, por ser diferencia de dos funciones derivables en dicho punto, siendo $EDN(f) = E'(c)$.

El siguiente teorema, cuya demostración es sencilla, relaciona la exactitud de una fórmula como la (5.1) con el hecho de ser de tipo interpolatorio.

Teorema 32 *La fórmula (5.1) es exacta para todo polinomio de grado menor o igual que* n *si y sólo si es de tipo interpolatorio (o sea si* $a_i = l_i'(c)$ *para* $i = 0, 1, \ldots, n$*, siendo* $l_i(x)$ *los polinomios de base de Lagrange).*

5.1.1. Fórmulas de derivación numérica de tipo interpolatorio y expresión del error

Veamos algunas de las fórmulas más usuales de derivación numérica de tipo interpolatorio, cuando tan sólo se conoce la función en uno, dos, tres o cuatro puntos, así como las expresiones de los errores correspondientes:

1. Si sólo se conoce el valor de f en un punto x_0, $f(x_0)$, entonces $f'(c) \simeq 0$.

2. Si se conocen $f(x_0)$ y $f(x_1)$ entonces el polinomio de interpolación es la recta dada por $p(x) = f(x_0) + f[x_0, x_1](x - x_0)$ y resulta la fórmula

$$\boxed{f'(c) \simeq f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}} \tag{5.3}$$

ahora, si se toman $x_0 = c$ y $x_1 = c + h$, la (5.3) se escribe en la forma

$$\boxed{f'(c) \simeq \frac{f(c+h) - f(c)}{h}} \tag{5.4}$$

Que se conoce como fórmula en diferencias progresivas. Ahora, si f es de clase $\mathbb{C}^{(2)}$ en el intervalo $[c, c+h]$(o $[c+h, c]$) el error puede expresarse en la forma

$$\boxed{EDN(f) = -\frac{h}{2} f''(\xi)} \quad (5.4^*)$$

con ξ intermedio entre c y $c+h$, para probarlo basta con utilizar el desarrollo de Taylor.

En cambio, si se toman puntos simétricos con respecto a c, o sea $x_0 = c-h$ y $x_1 = c+h$ la fórmula (5.3) queda en la forma

$$\boxed{f'(c) \simeq \frac{f(c+h)-f(c-h)}{2h}} \tag{5.5}$$

que se conoce como fórmula en diferencias centrales, suponiendo que f es de clase $\mathbb{C}^{(3)}$ en el intervalo $[c-h, c+h]$, se obtiene para el error, utilizando el desarrollo de Taylor, la expresión

$$\boxed{EDN(f) = -\frac{h^2}{6} f'''(\xi)} \; (5.5^*)$$

con $\xi \in [c-h, c+h]$.

3. Fórmulas con tres puntos: x_0, x_1, x_2. En este caso el polinomio de interpolación puede escribirse como

$$p(x) = f(x_0) + f[x_0, x_1](x-x_0) + f[x_0, x_1, x_2](x-x_0)(x-x_1)$$

entonces se obtiene

$$\boxed{f'(c) \simeq f[x_0, x_1] + f[x_0, x_1, x_2](2c - x_0 - x_1)} \tag{5.6}$$

en particular, tomando $x_0 = c, x_1 = c+h, x_2 = c+2h$ se tiene

$$\boxed{f'(c) \simeq \frac{-f(c+2h)+4f(c+h)-3f(c)}{2h}} \tag{5.7}$$

en tanto que si f es de clase $\mathbb{C}^{(3)}$ en un intervalo que contenga a los puntos $c, c+h, c+2h$, se obtiene para el error, utilizando el desarrollo de Taylor, la expresión

$$\boxed{EDN(f) = \frac{h^2}{3} f'''(\xi)} \; (5.7^*)$$

con ξ intermedio entre c y $c+2h$.

4. Fórmulas con cuatro puntos. En este caso lo más usual es tomar $x_0 = c-2h, x_1 = c-h, x_2 = c+h$ y $x_3 = c+2h$; obteniéndose la expresión

$$\boxed{f'(c) \simeq \frac{-f(c+2h)+8f(c+h)-8f(c-h)+f(c-2h)}{12h}} \tag{5.8}$$

y si f es de clase $\mathbb{C}^{(5)}$ en un intervalo que contenga a los puntos, el error correspondiente puede escribirse en la forma

$$\boxed{EDN(f) = \frac{h^4}{30} f^{(4)}(\xi)} \quad (5.8^*)$$

con ξ intermedio a los nodos de interpolación.

5.1.2. Fórmulas de derivación numérica de orden superior

Estas fórmulas se obtienen sin dificultad, por el mismo procedimiento, derivando k veces el polinomio de interpolación para hallar un valor aproximado de $f^{(k)}(c)$, si bien en este caso el número de puntos a utilizar debe ser mayor que el orden de derivación, pues en otro caso se obtendría $f^{(k)}(c) \simeq p^{(k)}(c) = 0$.

En el caso particular de que $k = 2$, utilizando tres puntos x_0, x_1, x_2, obtendremos la aproximación

$$\boxed{f''(c) \simeq 2f[x_0, x_1, x_2]} \qquad (5.9)$$

y usando puntos simétricos respecto al central $x_0 = c-h, x_1 = c$ y $x_2 = c+h$, resulta la fórmula en diferencias centrales

$$\boxed{f''(c) \simeq \frac{f(c+h) - 2f(c) + f(c-h)}{h^2}} \qquad (5.10)$$

además, si f es de clase $\mathbb{C}^{(4)}$ en algún intervalo que contenga a $c-h$ y $c+h$, se puede probar fácilmente, por el desarrollo de Taylor, que el error correspondiente adopta la forma

$$\boxed{EDN(f) = -\frac{h^2}{12} f^{(4)}(\xi)} \quad (5.10^*)$$

con ξ intermedio entre $c-h$ y $c+h$.

5.1.3. Estabilidad de las fórmulas de derivación numérica

La estabilidad de un método numérico mide como este responde frente a errores de redondeo o en los datos. Veamos la definición con más precisión para el caso que nos ocupa.

Definición 18 *Se dirá que una fórmula de derivación numérica en $n+1$ puntos, que escribimos ahora en la forma*

$$f'(c) \simeq \sum_{i=0}^{n} a_i^{(n)} f(x_i)$$

es ***estable*** *si* $\forall(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n)$ *existe una constante M, independiente de n, tal que*

$$\mid \sum_{i=0}^{n} a_i^{(n)} f(x_i) \mid \leq M \max_k \mid \varepsilon_k \mid$$

A este respecto conviene recordar el siguiente teorema, cuya demostración omitimos.

Teorema 33 *Las fórmulas de derivación numérica son inestables.*

5.2. Integración numérica

En muchas ocasiones es necesario disponer del valor de la integral de una función f en un intervalo $[a, b]$, pero no nos es posible calcularla por disponer tan sólo de una tabla de valores de la función en cuestión o bien por ser su expresión analítica inmanejable. Entonces, lo más usual es aproximarla por una combinación lineal de los valores de f en los puntos x_i, $i = 0, 1, \ldots, n$, en los que f es conocida, es decir

$$\boxed{\int_a^b f(x)dx \simeq \sum_{i=0}^{n} a_i f(x_i)} \tag{5.11}$$

al tomar este valor como aproximación de la integral se comete un error de integración numérica, que denotamos por $EIN(f)$ y está definido por medio de la fórmula

$$\boxed{\int_a^b f(x)dx = \sum_{i=0}^{n} a_i f(x_i) + EIN(f)}$$

Nuestro objetivo en este punto será el estudio de fórmulas de integración (o cuadratura) numérica de este tipo y de los errores correspondientes. Diremos que la fórmula (5.11) es **exacta para la función** ϕ si $EIN(\phi) = 0$, y se dirá **exacta de orden, al menos, p** si $EIN(x^i) = 0$ para $i = 0, 1, \ldots, p$ y de orden exactamente p si además es $EIN(x^{p+1}) \neq 0$. Asimismo, se dirá que es de **tipo interpolatorio** si se ha obtenido integrando el polinomio de interpolación de f; es decir si f es integrable en el intervalo $[a, b]$

y es $p(x) = \sum_{i=0}^{n} f_i l_i(x)$ el polinomio de interpolación de f, en los puntos x_i, $(i = 0, 1, \ldots, n)$ de dicho intervalo, expresado en la forma de Lagrange, entonces se tendrá

$$\boxed{\int_a^b f(x)dx \simeq \int_a^b p(x)dx = \sum_{i=0}^{n}(\int_a^b l_i(x)dx)f(x_i) + EIN(f)} \qquad (5.12)$$

Puesto que el error de interpolación se define por la igualdad

$$f(x) = p(x) + E(x)$$

si f es integrable en $[a, b]$, $E(x)$ también lo será, por ser diferencia de dos funciones integrables en dicho intervalo, siendo $EIN(f) = \int_a^b E(x)dx$.

El siguiente teorema, cuya demostración es sencilla, relaciona la exactitud de la fórmula (5.11) con el hecho de ser de tipo interpolatorio.

Teorema 34 *La fórmula (5.11) es exacta para todo polinomio de grado menor o igual que* n *si y sólo si es de tipo interpolatorio, es decir si* $a_i = \int_a^b l_i(x)dx$ *para* $i = 0, 1, \ldots, n$*, siendo* $l_i(x)$ *los polinomios de base de Lagrange.*

5.2.1. Fórmulas de tipo interpolatorio

Veamos algunas de las fórmulas más usuales de integración numérica de tipo interpolatorio, cuando tan sólo se conoce la función en uno o dos puntos, así como las expresiones de los errores correspondientes (algunas de las cuales serán demostradas como ejercicio):

1. Cuando sólo se conoce el valor de f en un punto x_0 de intervalo $[a, b]$ se puede aproximar la integral en la forma:

$$\boxed{\int_a^b f(x)dx \simeq f(x_0)(b - a)} \qquad (5.13)$$

en tanto que el error correspondiente puede escribirse como

$$\boxed{EIN(f) = \int_a^b f[x_0, x](x - x_0)dx} \qquad (5.13^*)$$

En particular, si $x_0 = a$ se tiene la **fórmula del rectángulo**

$$\boxed{\int_a^b f(x)dx \simeq f(a)(b - a)} \qquad (5.14)$$

y si $f \in \mathbb{C}^{(1)}([a,b])$ su error viene dado por

$$\boxed{EIN(f) = f'(\xi)\frac{(b-a)^2}{2}} \quad (5.14^*)$$

con $\xi \in (a,b)$.

En tanto que si $x_0 = \frac{a+b}{2}$ se tiene la **fórmula del punto medio**

$$\boxed{\int_a^b f(x)dx \simeq f(\frac{a+b}{2})(b-a)} \qquad (5.15)$$

que si $f \in \mathbb{C}^{(2)}([a,b])$ admite un error dado por la expresión

$$\boxed{EIN(f) = \frac{f''(\xi)}{24}(b-a)^3} \quad (5.15^*)$$

2. Cuando se conoce el valor de f en dos puntos x_0 y x_1, se tiene para el polinomio de interpolación de f en dichos puntos la expresión

$$p(x) = f(x_0) + f[x_0, x_1](x - x_0)$$

e integrando resulta

$$\int_a^b p(x)dx = f(x_0)(b-a) + f[x_0, x_1][\frac{(b-x_0)^2 - (a-x_0)^2}{2}]$$

en particular tomando $x_0 = a$ y $x_1 = b$ se obtiene la fórmula de cuadratura numérica

$$\boxed{\int_a^b f(x)dx \simeq \frac{1}{2}(b-a)[f(a) + f(b)]} \qquad (5.16)$$

conocida como **fórmula del trapecio**; ahora si $f \in \mathbb{C}^{(2)}([a,b])$ su error viene dado por

$$\boxed{EIN(f) = -f''(\xi)\frac{(b-a)^3}{12}} \quad (5.16^*)$$

con $\xi \in (a,b)$.

5.2.2. Fórmulas de Newton-Côtes simples

Se denomina así a las fórmulas de cuadratura numérica de tipo interpolatorio en que los **puntos de interpolación están igualmente espaciados** y resultan de la división en partes iguales del intervalo de integración $[a, b]$, si es $h = \frac{b-a}{n}$ entonces $x_i = x_0 + ih$ siendo $x_0 = a$. Si se utilizan los puntos extremos se dicen **cerradas** y en otro caso abiertas. Veamos seguidamente las fórmulas cerradas con tres, cuatro o cinco puntos:

1. Con tres puntos igualmente espaciados: a, $\frac{a+b}{2}$ y b, se calcula el polinomio de interpolación que pasa por ellos, se integra en $[a, b]$ y se obtiene la conocida **fórmula de Simpson** (una de las más usadas en la práctica por su simplicidad y precisión):

$$\boxed{\int_a^b f(x)dx \simeq \frac{b-a}{6}[f(a) + 4f(\frac{a+b}{2}) + f(b)]} \qquad (5.17)$$

que si $f \in \mathbb{C}^{(4)}([a, b])$ nos permite obtener la siguiente expresión del error

$$\boxed{EIN(f) = -f^{(4)}(\xi)\frac{(b-a)^5}{2880} = -f^{(4)}(\xi)\frac{h^5}{90}} \quad (5.17^*)$$

en la que $h = \frac{b-a}{2}$, en tanto que ξ es un punto intermedio entre a y b. De la expresión del error se deduce que es exacta para todo polinomio de grado menor o igual a 3.

2. Con cuatro puntos igualmente espaciados a, $\frac{2a+b}{3}$, $\frac{a+2b}{3}$ y b se obtiene

$$\boxed{\int_a^b f(x)dx \simeq \frac{b-a}{8}[f(a) + 3f(\frac{2a+b}{3}) + 3f(\frac{a+2b}{3}) + f(b)]} \quad (5.18)$$

pudiendo expresarse su error, si $f \in \mathbb{C}^{(4)}([a, b])$, en la forma

$$\boxed{EIN(f) = -f^{(4)}(\xi)\frac{3h^5}{80}} \quad (5.18^*)$$

siendo ahora $h = \frac{b-a}{3}$.

3. Con cinco puntos $x_0 = a, x_1 = a + h, x_2 = a + 2h, x_3 = a + 3h, x_4 = a + 4h = b$ (donde $h = \frac{b-a}{4}$) se obtiene la fórmula de cuadratura

$$\boxed{\int_a^b f(x)dx \simeq \frac{b-a}{90}(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4)} \qquad (5.19)$$

en tanto que el error puede escribirse como

$$\boxed{EIN(f) = -f^{(6)}(\xi)\frac{8h^7}{945}} \quad (5.19^*)$$

La mayor precisión del método de Simpson con $2+1$ puntos con respecto a sus rivales, con $1+1$ o $3+1$ puntos, es general; de hecho, se puede demostrar el teorema siguiente, que admitiremos sin demostración.

Teorema 35 *En las fórmulas de Newton-Côtes cerradas con n par ($n+1$ nodos), el error es un infinitésimo del orden de h^{n+3}, por contra, para n impar el error es del orden de h^{n+2} (De aquí la ventaja de utilizar fórmulas como la de Simpson, con n par).*

5.2.3. Fórmulas de cuadratura compuestas

Son las que se obtienen al dividir el intervalo de integración $[a, b]$ en n subintervalos iguales y a cada uno de estos aplicarles una fórmula sencilla, por ejemplo la fórmula del trapecio o la regla de Simpson, estas fórmulas tienen buenas propiedades de estabilidad y convergencia. Veamos a modo de ejemplo las fórmulas compuestas del trapecio y de Simpson.

1. Fórmula del **trapecio compuesta**. Dividamos el intervalo de partida $[a, b]$ en n subintervalos de la misma amplitud h, siendo esta $h = \frac{b-a}{n}$, sean $x_0 = a, x_j = x_0 + jh$ y apliquemos a cada subintervalo la fórmula del trapecio simple, entonces resulta

$$\int_a^b f(x)dx = \sum_{j=0}^{n-1}\int_{x_j}^{x_{j+1}} f(x)dx \simeq \sum_{j=0}^{n-1}\frac{h}{2}[f(x_j) + f(x_{j+1})]$$

de donde se obtiene

$$\boxed{\int_a^b f(x)dx \simeq \frac{h}{2}[f_0 + 2\sum_{j=1}^{n-1} f_j + f_n]} \quad (5.20)$$

en la que $f_j = f(x_j)$; ahora si $f \in \mathbb{C}^{(2)}([a, b])$ el error puede escribirse, tras sumar los errores cometidos en cada subintervalo, en la forma

$$\boxed{EIN(f) = -f''(\xi)\frac{(b-a)^3}{12n^2}} \quad (5.20^*)$$

siendo ξ un punto intermedio entre a y b.

2. Fórmula de **Simpson compuesta**. De la misma manera que antes dividamos el intervalo de partida en n subintervalos de la misma longitud, suponiendo que f es conocida en todos los extremos x_j de los subintervalos y en los puntos medios de los mismos que indicamos por $x_{j+\frac{1}{2}} = x_j + \frac{h}{2}$, aplicando a cada subintervalo la regla de Simpson simple se obtiene la fórmula de Simpson compuesta

$$\boxed{\int_a^b f(x)dx \simeq \frac{h}{6}[f_0 + 2\sum_{j=1}^{n-1} f_j + 4\sum_{j=0}^{n-1} f_{j+\frac{1}{2}} + f_n]} \qquad (5.21)$$

y si $f \in \mathbb{C}^{(4)}([a,b])$, sumando los errores cometidos en cada subintervalo se obtiene el error de la fórmula compuesta, dado por la fórmula

$$\boxed{EIN(f) = -f^{(4)}(\xi)\frac{(b-a)^5}{2880n^4}} \; (5.21^*)$$

con ξ intermedio entre a y b.

5.2.4. Estabilidad y convergencia

Como se ha dicho anteriormente, la estabilidad de un método numérico mide su sensibilidad frente a errores de cálculo o en los datos. Para el caso de las fórmulas de cuadratura se da la siguiente.

Definición 19 *Se dirá que una fórmula de integración numérica en $n+1$ puntos, que escribimos ahora en la forma*

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{n} a_i^{(n)} f(x_i)$$

*es **estable** si $\forall(\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n)$ existe una constante M, independiente de n, tal que:*

$$\mid \sum_{i=0}^{n} a_i^{(n)} f(x_i) \mid \leq M \max_k \mid \varepsilon_k \mid$$

También se introduce el concepto de convergencia como sigue.

Definición 20 *Una fórmula de integración numérica en $n+1$ puntos*

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{n} a_i^{(n)} f(x_i)$$

se dirá ***convergente*** *sobre un conjunto* V *si cualquiera que sea* $f \in V$ *se tiene*

$$lim_{n\to\infty} \mid \sum_{i=0}^{n} a_i^{(n)} f(x_i) \mid = \int_a^b f(x)dx$$

Notas. Damos seguidamente algunos resultados sobre estabilidad y convergencia.

- Las fórmulas de Newton-Côtes simples son inestables (luego no son recomendables para valores grandes de n).
- Una condición necesaria y suficiente para que un método de cuadratura de tipo interpolatorio sea convergente sobre $\mathbb{C}([a,b])$ es que sea estable.
- Las fórmulas del trapecio y Simpson compuestas son estables y convergentes; en general, las fórmulas compuestas de Newton-Côtes, de grado de exactitud r en $n+1$ puntos, tales que los coeficientes de la correspondiente fórmula simple son positivos son estables.

5.3. Problemas resueltos

1. Obtener la fórmula de tipo interpolatorio para aproximar la $f''(a)$ si se conoce f en $a, a+h, a+2h$.

   Solución. Hemos de aproximar $f''(a) \cong p''(a)$ siendo $p(x)$ el polinomio que interpola a f en los puntos $a, a+h, a+2h$, que puede escribirse en la forma

   $$p(x) = f[a] + f[a, a+h](x-a) + f[a, a+h, a+2h](x-a)(x-a-h)$$

   por tanto $p''(x) = 2f[a, a+h, a+2h]$ para todo x, en consecuencia

   $$f''(a) \cong 2f[a, a+h, a+2h] = 2\frac{f[a+h, a+2h] - f[a, a+h]}{2h} =$$

   $$= \frac{\frac{f[a+2h]-f[a+h]}{h} - \frac{f[a+h]-f[a]}{h}}{h} = \boxed{\frac{f(a+2h)-2f(a+h)+f(a)}{h^2}}$$

2. Deducir una fórmula de derivación numérica de tipo interpolatorio para aproximar $f''(c)$ utilizando los valores de f en $c-h$, c y $c+h$ y obtener la expresión del error correspondiente suponiendo que f es de clase $\mathcal{C}^{(4)}$ en un intervalo conteniendo a dichos puntos.

Solución. De un modo similar al ejercicio anterior obtenemos la aproximación de la $f''(c)$ que sigue

$$f''(c) \cong \frac{f(c+h) - 2f(c) + f(c-h)}{h^2}$$

y suponiendo que f es de clase $\mathcal{C}^{(4)}$ en el intervalo $[c-h, c+h]$, podemos escribir los siguientes desarrollos de Taylor

$$f(c+h) = f(c)+f'(c)h+\frac{f''(c)}{2!}h^2+\frac{f'''(c)}{3!}h^3+\frac{f^{(iv)}(\xi_1)}{4!}h^4 \text{ con } \xi_1 \in (c, c+h)$$

$$f(c-h) = f(c)-f'(c)h+\frac{f''(c)}{2!}h^2-\frac{f'''(c)}{3!}h^3+\frac{f^{(iv)}(\xi_2)}{4!}h^4 \text{ con } \xi_2 \in (c-h, c)$$

entonces, sumando ambas, pasando $2f(c)$ al primer miembro y dividiendo por h^2 se obtiene

$$\frac{f(c+h) - 2f(c) + f(c-h)}{h^2} = f''(c) + (f^{(iv)}(\xi_1) + f^{(iv)}(\xi_2))\frac{h^2}{4!}$$

y puesto que

$$\frac{f^{(iv)}(\xi_1) + f^{(iv)}(\xi_2)}{2} = f^{(iv)}(\xi) \text{ con } \xi \in (c-h, c+h)$$

ya que al ser la derivada $f^{(iv)}(x)$ continua, se alcanza el valor medio entre dos cualesquiera de la misma en un punto intermedio ξ entre ξ_1 y ξ_2, se tiene

$$f''(c) = \frac{f(c+h) - 2f(c) + f(c-h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi)$$

por tanto, el error de esta fórmula de derivación numérica es

$$EDN(f) = -\frac{h^2}{12}f^{(iv)}(\xi)$$

3. Hallar el número de subintervalos que deben tomarse para aproximar el valor de la integral $\int_1^2 cos\sqrt{x}dx$, mediante la regla del trapecio compuesta, de manera que el módulo del error cometido sea menor que 0,005.

   Solución. El error de la fórmula del trapecio compuesta viene dado por la expresión

$$EIN(f) = -f''(\xi)\frac{(b-a)^3}{12n^2}$$

siendo a y b los extremos del intervalo de integración, en nuestro caso $a = 1$ y $b = 2$, n el número de partes iguales en las que se subdivide dicho intervalo y ξ un punto intermedio desconocido. Luego hemos de hallar n para que $\mid EIN(f) \mid$ cumpla dicha condición, o sea tal que

$$\mid \frac{f''(\xi)}{12n^2} \mid < 5 \cdot 10^{-3}$$

como $f(x) = \cos\sqrt{x}$ es $f'(x) = -\frac{\sin\sqrt{x}}{2\sqrt{x}}$ y su derivada segunda viene dada por

$$f''(x) = \frac{-\sqrt{x}\cos\sqrt{x} + \sin\sqrt{x}}{4x\sqrt{x}}$$

por tanto para $\xi \in [1, 2]$, será

$$\mid f''(\xi) \mid = \mid \frac{-\sqrt{\xi}\cos\sqrt{\xi} + \sin\sqrt{\xi}}{4\xi\sqrt{\xi}} \mid < \frac{\sqrt{\xi}+1}{4\xi\sqrt{\xi}} < \frac{3}{4}$$

ya que los módulos de seno y coseno de $\xi \in [1, 2]$ son, en todo caso, menores que 1 y por otro lado, en la última fracción el numerador es menor que $2 + 1$ y el denominador mayor que 4 para todo $\xi \in [1, 2]$. Por tanto, es suficiente con tomar n verificando la desigualdad

$$\mid \frac{f''(\xi)}{12n^2} \mid < \frac{\frac{3}{4}}{12n^2} = \frac{1}{16n^2} < 5 \cdot 10^{-3}$$

o lo que es equivalente

$$n^2 > \frac{1000}{5 \cdot 16} = 12{,}5 \Leftrightarrow n > 3{,}5355...$$

basta pues con subdividir el intervalo $[1, 2]$ en 4 partes iguales para asegurarnos que la regla del trapecio compuesta nos da el resultado con la aproximación requerida.

4. Aproximar la integral $\int_0^1 x\cos(x)dx$, por la regla de Simpson compuesta, de manera que el módulo del error cometido sea menor que $5 \cdot 10^{-5}$ (justificar la respuesta).

Solución. Recordemos que el error de la fórmula de Simpson compuesta viene dado por la expresión

$$EIN(f) = -f^{(iv)}(\xi)\frac{(b-a)^5}{2880n^4}$$

siendo a y b los extremos del intervalo de integración, en este caso $a = 0$ y $b = 1$, n el número de partes iguales en las que se subdivide dicho

intervalo y ξ un punto intermedio entre 0 y 1. Luego es suficiente con de hallar n tal que

$$\mid \frac{f^{(iv)}(\xi)}{2880n^4} \mid < 5 \cdot 10^{-5}$$

ahora fácilmente se deduce que

$$f^{(iv)}(x) = 4\sin(x) + x\cos(x)$$

por tanto $\mid f^{(iv)}(\xi) \mid < 5$, pues los módulos de las funciones seno y coseno son menores que 1 y la ξ está entre 0 y 1, en consecuencia es suficiente con tomar n de manera que se verifique la desigualdad

$$5/(2880n^4) < 5 \cdot 10^{-5} \Longleftrightarrow n^4 > 100000/2880$$

$$\Longleftrightarrow n > \sqrt[4]{100000/2880} = 2{,}427458858536617$$

tomemos pues n = 3, entonces, como h = 1/3, tendremos el siguiente valor aproximado de la integral:

$$\int_0^1 x\cos(x)dx \cong (1/18)[f(0) + f(1) + 2(f(1/3) + f(2/3)) +$$

$$+4(f(1/6) + f(0{,}5) + f(5/6))] = 0{,}38178285...$$

que difiere del valor exacto 0,38177329... menos de lo requerido.

5. Determinar en cuántos subintervalos de igual longitud hay que subdividir el intervalo $[1, 2]$, para calcular $\int_1^2 \frac{1}{x}dx$ por las reglas compuestas del trapecio y de Simpson con módulo del error menor que $5 \cdot 10^{-4}$, con la que se requiera menor número calcular el valor aproximado de $\log 2$.

   Solución. En primer lugar, es inmediato ver que $\int_1^2 \frac{1}{x}dx = \log(x) \mid_1^2 = \log(2)$ (donde $\log(x)$ es el logaritmo neperiano o natural de x). Ahora, recordemos que el error de la fórmula del trapecio compuesta viene dado por la expresión

$$EIN(f) = -f''(\xi)\frac{(b-a)^3}{12n^2}$$

   siendo a y b los extremos del intervalo de integración, en nuestro caso $a = 1$ y $b = 2$, n el número de partes iguales en las que se subdivide dicho intervalo y ξ un punto intermedio desconocido. Luego hemos de hallar n tal que

$$\mid \frac{f''(\xi)}{12n^2} \mid < 5 \cdot 10^{-4}$$

en este caso $| f''(\xi) |= f''(\xi) = \frac{2}{\xi^3}$ y toma su valor máximo para ξ entre 1 y 2, cuando ξ, que está en el denominador, es menor o sea para $\xi = 1$, resultando máx $| f''(\xi) |< 2$, por tanto para asegurar la aproximación de la integral requerida por este método es suficiente que n verifique la relación

$$\frac{2}{12n^2} < 5 \cdot 10^{-4} \Longleftrightarrow n^2 > \frac{10000}{30} \Longleftrightarrow n > 18{,}2574...$$

luego tomando $n = 19$ o mayor aseguramos la aproximación requerida por la regla del trapecio compuesta de la integral, es decir del $\log(2)$.

Para la regla de Simpson compuesta, a la vista de la expresión del error correspondiente, es suficiente con tomar n verificando

$$| \frac{f^{(iv)}(\xi)}{2880n^4} |< 5 \cdot 10^{-4}$$

puesto que $| f^{(iv)}(\xi) |= f^{(iv)}(\xi) = \frac{4!}{\xi^5}$ toma su valor máximo para ξ entre 1 y 2, cuando ξ, que está en el denominador, es menor o sea para $\xi = 1$, resultando máx $| f^{(iv)}(\xi) |< 24$, por tanto para asegurar la aproximación de la integral requerida por este método es suficiente que n verifique la relación

$$\frac{24}{2880n^4} < 5 \cdot 10^{-4} \Longleftrightarrow n^4 > \frac{240000}{5 \cdot 2880} \Longleftrightarrow n > 2{,}0205...$$

luego lo haremos por la regla de Simpson compuesta con $n = 3$, que requiere muchos menos cálculos, resultando

$$\log(2) = \int_1^2 \frac{1}{x}dx \cong \frac{1}{18}[f(1) + f(2) + 2(f(4/3) + f(5/3))+$$

$$+4(f(7/6) + f(1{,}5) + f(11/6))] = 0{,}69316979...$$

que difiere del valor exacto de $\log(2) = 0{,}69314718055995...$ menos de lo requerido.

6. Determinar en cuántos subintervalos de igual longitud hay que subdividir el intervalo $[0, 1]$, para calcular $\int_0^1 e^{-x^2}dx$ por las reglas compuestas del trapecio y de Simpson con error menor que $5 \cdot 10^{-5}$, con la que se requiera menor número calcular el valor aproximado de dicha integral.

Solución. Ahora es $a = 0$, $b = 1$, $f(x) = e^{-x^2}$ y nos requieren para la regla del trapecio compuesta que

$$| \frac{f''(\xi)}{12n^2} |< 5 \cdot 10^{-5}$$

pero $f'(x) = -2xe^{-x^2}$ y $f''(x) = -2e^{-x^2} + 4x^2e^{-x^2} = e^{-x^2}(4x^2-2)$ como $\mid f''(\xi) \mid \leq \text{máx}_{x\in[0,1]} \mid f''(x) \mid = 2$ (para probar esto, basta con tener en cuenta que el máximo de una función continua en un intervalo cerrado y acotado, en este caso el $[0,1]$, se alcanza en alguno de los extremos o de los puntos críticos interiores, y como no existen basta ver donde es mayor si en $x = 0$ o en $x = 1$). Por tanto, para asegurar la aproximación requerida es suficiente con elegir n verificando la desigualdad

$$\frac{2}{12n^2} 5 \cdot 10^{-5} \Longleftrightarrow n^2 > \frac{100000}{30} \Longleftrightarrow n > 57{,}73502691896...$$

En cambio para la regla se Simpson compuesta n debería verificar la desigualdad

$$\mid \frac{f^{(iv)}(\xi)}{2880n^4} \mid < 5 \cdot 10^{-5}$$

las derivadas tercera y cuarta de f están dadas por $f'''(x) = e^{-x^2}(-8x^3 + 12x)$ y $f^{(iv)}(x) = e^{-x^2}(16x^4 - 48x^2 + 12)$ y el módulo de esta última alcanza su máximo valor, 12, en el extremo $x = 0$ (es aplicable el razonamiento entre paréntesis anterior). Así, para asegurar la aproximación requerida es suficiente con elegir n verificando la desigualdad

$$\mid \frac{12}{2880n^4} \mid < 5 \cdot 10^{-5} \Longleftrightarrow n^4 > \frac{1200000}{5 \cdot 2880} \Longleftrightarrow n > 3{,}021375397...$$

Calculemos pues el valor aproximado de dicha integral tomando $n = 4$, es decir $h = 1/4 = 0{,}25$, entonces se tendrá

$$\int_0^1 e^{-x^2} dx \cong \frac{1}{24}\{f(0)+f(1)+2[f(0{,}25)+f(0{,}5)+f(0{,}75)]+4[f(0{,}125)+$$

$$+f(0{,}375) + f(0{,}625) + f(0{,}875)]\} = 0{,}74682612...$$

que aproxima el valor exacto 0,74682413... con error menor que el requerido.

5.4. Algunas funciones Maxima para la integración numérica

5.4.1. Regla del trapecio compuesta

Construiremos una función de Maxima, mediante un block, para el cálculo aproximado de integrales mediante la regla del trapecio compuesta; para ello,

consideramos la función f en el intervalo $[a,b]$, definimos una partición del intervalo de amplitud $h=(b-a)/n$, dividiendo el intervalo en n partes iguales y aplicando la fórmula del trapecio simple a cada subintervalo $[x_i,x_{i+1}]$ para $i=0,1,...,n-1$, resulta

```
(%i1)  TC(f,a,b,n):=block([numer],numer:true,
       E:f(a)+f(b),h:(b-a)/n,for i:1 thru n-1 do(x[i]:a+i*h),
       I:0,for i:1 thru n-1 do (I:I+f(x[i])),
       print('integrate(f(x),x,a,b),"(con TC)~",h*(E/2+I)))$
```

Seguidamente la aplicamos para aproximar la integral de $f(x)=cos(x^2)$ en el intervalo $[0,1]$.

```
(%i2)  f(x):=cos(x^2)$ TC(f,0,1,100)$
```

$\int_0^1 \cos(x^2)dx$(con TC) $\simeq$ 0,90451021338042

Dado que conocemos la expresión del error de la fórmula del trapecio compuesta, caso de ser la función a integrar de clase $\mathcal{C}^{(2)}$ y de que se pueda acotar el valor absoluto de $f''(x)$ en el intervalo dado, puede mejorarse el programa, obteniendo previamente el número n de subdivisiones a realizar para conseguir que el módulo del error absoluto cometido sea menor que un épsilon dado.

5.4.2. Regla de Simpson compuesta

Definiremos otra función de Maxima, mediante un block, para el cálculo aproximado de integrales mediante la regla de Simpson compuesta; para ello, consideramos la función f en el intervalo $[a,b]$, definimos una partición del intervalo de amplitud $h=(b-a)/2n$, dividiendo el intervalo en $2n$ partes iguales y aplicamos la fórmula de Simpson simple n veces, una a cada subintervalo $[x_i,x_{i+2}]$ para $i=0,2,...,2(n-1)$, siendo $x_0=a$, $x_i=x_0+ih$ $(i=1,2,...,2n)$ y $x_{2n}=b$; puesto que la amplitud de cada uno de los n subintervalos es $2h$ resulta que $2h/6=h/3$, por tanto la integral de f en cada subintervalo se aproxima por $(h/3)[f(x_i)+4f(x_{i+1})+f(x_{i+2}]$; entonces, denotando la contribución de los puntos extremos por E, la de los puntos intermedios de índice impar por I_i (estos son los puntos medios de los intervalos donde se aplica la regla de Simpson simple) y por I_p la correspondiente a los puntos intermedios de índice par (estos son los puntos que separan los subintervalos donde se aplica la fórmula de Simpson simple), resultará la fórmula de Simpson compuesta dada por la siguiente función.

```
(%i4)  SC(f,a,b,n):=block([numer],numer:true,
       E:f(a)+f(b),h:(b-a)/(2*n),for i:0 thru 2*n
       do(x[i]:a+i*h),
       I_i:0,for i:1 step 2 thru 2*n-1 do (I_i:I_i+f(x[i])),
       I_p:0, for i:2 step 2 thru 2*n-2 do (I_p:I_p+f(x[i])),
       print('integrate(f(x),x,a,b),"(con SC)~",
       (h/3)*(E+4*I_i+2*I_p)))$
```

Aplicando dicha regla a la integral anterior se tendrá

```
(%i5)  f(x):=cos(x^2)$ SC(f,0,1,100)$
```

$\int_0^1 \cos(x^2)dx$(con SC) $\simeq$ 0,90452423790113

Al igual que en el caso anterior, dado que conocemos la expresión del error de la fórmula de Simpson compuesta, caso de ser la función a integrar de clase $\mathcal{C}^{(4)}$ y de que se pueda acotar el valor absoluto de $f^{iv}(x)$ en el intervalo dado, puede mejorarse el programa, obteniendo previamente el número n de subdivisiones a realizar para conseguir que el módulo del error absoluto cometido sea menor que un épsilon prefijado. Ahora apliquemos ambas reglas a la función $f(x) = 1/(5-x)$ en el intervalo $[2,3]$, y comparemos los resultados con el valor exacto de la misma dado por $log(3/2) = 0{,}40546510810816$

```
(%i7)  f(x):=1/(5-x);
       'integrate(f(x),x,2,3)= float(integrate(f(x),x,2,3));
       print("Tomamos el número de pasos n = ",n:100)$
       print("Módulo del error de la regla TC con ",n," pasos es ",
       abs(TC(f,2,3,n)-log(1.5)))$
       print("Módulo del error de la regla SC con ",n," pasos es ",
       abs(SC(f,2,3,n)-log(1.5)))$
```

(%o7) $\mathrm{f}(x) := \dfrac{1}{5-x}$

(%o8) $\int_2^3 \frac{1}{5-x}dx = 0{,}40546510810816$

Tomamos el número de pasos $n = 100$

$\int_2^3 \frac{1}{5-x}dx$(con TC) $\simeq$ 0,40546626551139

Módulo del error de la regla TC con 100 pasos es $1{,}157403227758369 \cdot 10^{-6}$

$\int_2^3 \frac{1}{5-x}dx$(con SC) $\simeq$ 0,40546510810921

Módulo del error de la regla SC con 100 pasos es $1{,}0447753773235036 \cdot 10^{-12}$

5.5. Problemas y trabajos propuestos

Problemas propuestos:

1. Hallar a_1 y a_2 para que las fórmulas de derivación e integración numéricas:
$$f'(\frac{1}{2}) \simeq a_1 f(0) + a_2 f(\frac{1}{2})$$
y
$$\int_0^1 f(x)dx \simeq a_1 f(0) + a_2 f(\frac{1}{2})$$
sean exactas para las funciones 1 y x.

2. Obtener una fórmula de derivación numérica de tipo interpolatorio para el cálculo aproximado de $f'(a)$, siendo conocidos los valores de f en los puntos $a, a+\frac{h}{4}, a+\frac{h}{2}$ y $a+h$. Utilizar dicha fórmula par aproximar la derivada de $f(x) = cos(hx^2)$ con $x = a = 1$ y $h = 0{,}1$.

3. Calcular a_1, a_2 y a_3 para que la fórmula
$$\int_{-1}^1 f(x)dx \simeq a_1 f(-1) + a_2 f(1) + a_3 f(\alpha)$$
sea exacta del mayor grado posible, siendo α un número dado tal que $-1 < \alpha < 1$. Utilizar dicha fórmula para estimar el valor de la integral
$$\int_{-1}^1 \sqrt{\frac{5x+13}{2}}dx$$
con $\alpha = -0{,}1$. Comparar el valor obtenido con el verdadero valor de dicha integral.

4. Supuestos conocidos los valores de f en los puntos $c-2h, c-h, c+h, c+2h$; hallar una fórmula para aproximar la derivada $f'''(c)$ y dar una expresión del error si f es de clase $\mathbb{C}^{(5)}$ en un intervalo que contenga a dichos puntos.

5. Calcular los coeficientes y nodos para que la fórmula de integración numérica $\int_{-1}^1 f(x)dx \simeq A_0 f(x_0) + A_1 f(x_1)$ tenga el máximo orden de exactitud. Aplicarla para calcular aproximadamente la integral
$$\int_{-1}^1 (1+x^2)senx dx$$

6. Aplicar la regla de Simpson compuesta a la integral $\int_1^x \frac{1}{t}dt$, para obtener una aproximación de $\log 2$, determinando el número de subintervalos a considerar para que el error cometido en esa aproximación sea menor que 10^{-3}.

7. Utilizar la regla del trapecio compuesta para aproximar la integral $\int_{1,8}^{3,4} e^x dx$, eligiendo el paso h para que el error cometido sea menor que $5 \cdot 10^{-3}$. Con el mismo paso anterior, integrarla por la regla de Simpson compuesta y estimar el error correspondiente.

8. Obtener el valor aproximado de la integral $\int_1^2 (lnx)^2 dx$ utilizando las fórmulas compuestas del punto medio, trapecio y Simpson, dividiendo el intervalo en ocho subintervalos y estimando el error en cada caso.

Trabajos propuestos:
Se propone en este tema realizar alguno de los siguientes trabajos:

- Estabilidad y convergencia en los problemas de derivación e integración numérica.
- Fórmulas de cuadratura gaussianas.
- Método de Romberg para derivación e integración numérica.

Capítulo 6

Resolución numérica de problemas de valor inicial

6.1. Problemas de valor inicial para ecuaciones diferenciales ordinarias

Se denomina problema de valor inicial, abreviadamente PVI, para la ecuación diferencial ordinaria $y'(t) = f(t, y(t))$ al problema de hallar una solución de la misma verificando una condición inicial dada $y(a) = \eta$. Es decir, se trata de obtener una función $y(t)$, definida en algún intervalo I conteniendo al punto a, verificando las condiciones

$$PVI \begin{cases} y'(t) = f(t, y(t)), \forall t \in I \\ y(a) = \eta \end{cases}$$

En primer lugar, enunciamos el siguiente teorema que garantiza la existencia y unicidad de solución del problema de valor inicial planteado anteriormente, en el caso escalar.

Teorema 36 *Sea $f : D = [a, b] \times \mathbb{R} \to \mathbb{R}$, con a y b reales finitos, una función continua verificando una condición de Lipschitz uniforme, con respecto a y en D, de la forma: existe L constante positiva tal que cualesquiera que sean $t, y, \widetilde{y}$ con $(t, y), (t, \widetilde{y}) \in D$ es*

$$|f(t, y) - f(t, \widetilde{y})| \leq L \mid y - \widetilde{y} \mid$$

Entonces, si η es cualquier número real dado, existe una única solución $y(t)$ del PVI que se denota por $y(t; a, \eta)$, que es de clase $C^{(1)}$ con respecto a $t \in [a, b]$ y continua con respecto a η.

Observación: Damos aquí una condición suficiente para que $f(t,y)$ verifique una condición de Lipschitz uniforme del tipo anterior.

Si $f(t,y)$ y $\frac{\partial f(t,y)}{\partial y}$ son continuas para todo $(t,y) \in D$, siendo D convexo (lo es en las hipótesis del teorema anterior), por el teorema del valor medio se tiene

$$f(t,y) - f(t,\widetilde{y}) = \frac{\partial f(t,\widehat{y})}{\partial y}(y - \widetilde{y})$$

con $y < \widehat{y} < \widetilde{y}$ o bien $\widetilde{y} < \widehat{y} < y$. Entonces, si existe $\sup_{(t,y)\in D} \mid \frac{\partial f(t,y)}{\partial y} \mid$, $f(t,y)$ verificará la condición de Lipschitz anterior, pues bastaría con tomar L igual a cualquier positivo mayor o igual que dicho supremo.

Para el caso vectorial tenemos análogamente el siguiente teorema

Teorema 37 *Sea $f : D = [a,b] \times \mathbb{R}^s \to \mathbb{R}^s$, con a y b finitos, una función continua verificando la siguiente condición de Lipschitz uniforme con respecto a y en D: existe L constante positiva de manera que cualesquiera que sean $t, y, \widetilde{y}$ con $(t,y), (t,\widetilde{y}) \in D$ es*

$$\| f(t,y) - f(t,\widetilde{y}) \| \leq L \| y - \widetilde{y} \|$$

Entonces, si η es cualquier vector dado de $\mathbb{R}^s$, existe una única solución vectorial $y(t)$ del PVI que se denota por $y(t;a,\eta)$, que es de clase $C^{(1)}$ con respecto a $t \in [a,b]$ y continua con respecto a $\eta \in \mathbb{R}^s$. Si se sustituye la condición inicial $y(a) = \eta$ por $y(t_0) = y_0$ con $t_0 \in [a,b]$ e $y_0 \in \mathbb{R}^s$, también existe una única solución $y(t;t_0,y_0)$, de clase $C^{(1)}$ con respecto a $t \in [a,b]$, que depende continuamente de los datos (t_0,y_0). Si además de las condiciones anteriores fuese $f \in C^{(p)}$ en D, entonces la solución $y(t;t_0,y_0)$ sería de clase $C^{(p+1)}$ respecto a t en $[a,b]$ y de clase $C^{(p)}$ con respecto a y_0 en $\mathbb{R}^s$.

Observaciones:

- Aquí $\| \cdot \|$ es una norma cualquiera en $\mathbb{R}^s$ (se indicará indistintamente por $| \cdot |$ si no hay lugar a confusión).
- Como antes, si $f(t,y)$ y $\partial f(t,y)/\partial y$ (Jacobiana de f respecto a y) son continuas y existe $\sup_{(t,y)\in D} \| \partial f(t,y)/\partial y \|$, la función $f(t,y)$ verificará la condición de Lipschitz anterior.
- El problema de valor inicial vectorial se escribe desarrollado en la forma:

$$\begin{array}{ll} y_1' = f_1(t,y_1,y_2,...,y_s), & y_1(a) = \eta_1 \\ y_2' = f_2(t,y_1,y_2,...,y_s), & y_2(a) = \eta_2 \\ \multicolumn{2}{c}{..............} \\ y_s' = f_s(t,y_1,y_2,...,y_s), & y_s(a) = \eta_s \end{array}$$

- En lo sucesivo indicaremos por $C_L = \{f : D = [a,b] \times \mathbb{R}^s \to \mathbb{R}^s \mid f$ es continua y verifica una condición de Lipschitz del tipo anterior$\}$ y por $C^{(p)} = \{f : D = [a,b] \times \mathbb{R}^s \to \mathbb{R}^s \mid$ tal que f admite todas las derivadas parciales hasta el orden p y son continuas en $D\}$.

Ecuaciones de orden superior al primero

En la práctica suelen aparecer problemas de valor inicial para ecuaciones diferenciales ordinarias de orden $s > 1$, que escritas en la forma normal o explícita se expresan como sigue

$$y^{(s)} = f(t, y, y', y'', ..., y^{(s-1)}), \quad y^{(k)}(a) = \eta_k (k = 0, 1, 2, ..., s-1)$$

Entonces, definiendo las funciones vectoriales

$$y = \begin{pmatrix} y_1 & \equiv & y \\ y_2 & \equiv & y' \\ \vdots & \vdots & \vdots \\ y_s & \equiv & y^{(s-1)} \end{pmatrix}, \; F(t,y) = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ f(t, y_1, y_2, ..., y_s) \end{pmatrix}, \; \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_s \end{pmatrix}$$

El problema del valor inicial dado será equivalente al PVI para el sistema de primer orden

$$y' = F(t, y), \quad y(a) = \eta$$

La equivalencia significa que si se tiene la solución del primero se pueden construir las del segundo y recíprocamente. El interés de este cambio radica en que, en general, nos referiremos a métodos numéricos para problemas de valor inicial escalares o vectoriales de primer orden, conviene recordar este modo simple de pasar de una ecuación de orden superior al primero a un sistema de primer orden equivalente. Este procedimiento puede generalizarse, sin dificultad, para pasar de varias ecuaciones de orden superior a un sistema equivalente de primer orden.

6.2. Métodos de un paso generales: definiciones y resultados

Cuando se estudia la resolución numérica de un problema de valor inicial es habitual suponer que está matemáticamente bien planteado, es decir, que tiene solución única y que esta depende continuamente de los datos iniciales. Con las hipótesis antes señaladas esto está garantizado. Aunque no es

suficiente para asegurar una buena resolución numérica; en efecto, si consideramos dos soluciones que salen en el instante $t = 0$ de y_0 e $\widehat{y}_0$ respectivamente, se puede probar que

$$\forall t \in [a, b] : | y(t, t_0, y_0) - y(t, t_0, \widehat{y}_0 | \leq e^{Lt} | y_0 - \widehat{y}_0 |$$

entonces, si el factor e^{Lt} se hace grande, una resolución numérica puede no tener sentido con medios de cálculo limitados, ya que los posibles errores de redondeo propagados podrían ser amplificados por el factor e^{Lt}. De manera que, en lo que sigue, consideraremos la resolución numérica de PVI que además de matemáticamente bien planteados estén numéricamente bien planteados en el sentido de que e^{Lt} no sea demasiado grande, donde esta terminología heurística depende de la precisión del ordenador o medio de cálculo utilizado. En la resolución numérica de un PVI de la forma

$$\begin{cases} y'(t) = f(t, y(t)), t \in [a, b] \\ y(a) = \eta \end{cases}$$

donde f es, indistintamente, escalar o vectorial, se suele considerar un conjunto de puntos en [a,b] de la forma

$$a = t_0 < t_1 < t_2 < .... < t_N = b$$

que llamamos red en $[a, b]$, y a los $h_j = t_{j+1} - t_j$, pasos de la red. En general, por simplicidad en el planteamiento, tomaremos todos los pasos $h_j = h$ fijos, aunque se podrían tomar variables siempre que el mayor de ellos tendiera a cero. Llamaremos **método de un paso** a todo algoritmo en el que a partir de los datos t_n, y_n (aproximación de la solución en el punto t_n de la red) y el paso h, podemos calcular un valor y_{n+1} que aproxima a la solución $y(t; t_n, y_n)$ del PVI: $y' = f(t, y), y(t_n) = y_n$ en el punto $t = t_n + h$.

Algunos de los métodos más antiguos de aproximación numérica de las soluciones de una ecuación diferencial están motivados por sencillas consideraciones geométricas basadas sobre el campo de direcciones (o pendientes) definidas por el miembro de la derecha de la ecuación diferencial; entre estos están los métodos de Euler y de Euler modificado. Otros más precisos y sofisticados están basados en los desarrollos en serie de Taylor y los veremos en un párrafo posterior.

El método de Euler fue propuesto por dicho autor en 1768, en los primeros días del cálculo. Consiste básicamente en seguir la pendiente del punto (t_n, y_n) sobre un intervalo de longitud h, así la aproximación en el punto $t_n + h$ está dada por

$$y_{n+1} = y_n + hf(t_n, y_n), (n = 0, 1, ..., N - 1)$$

y cuyo significado geométrico es bastante sencillo, consiste en dar como aproximación de la solución en el punto $t_{n+1} = t_n + h$, la que daría la recta tangente en el punto (t_n, y_n) a la curva integral del PVI: $y'(t) = f(t, y(t)), y(t_n) = y_n$, evaluada en el punto $t_{n+1} = t_n + h$.

6.2.1. Expresión general de los métodos de un paso

Según hemos dicho antes, un **método de un paso** es un algoritmo que relaciona la aproximación y_{n+1} a $y(t_{n+1}; t_n, y_n)$ con (t_n, y_n) y la longitud del paso h, se suele expresar como sigue

$$\begin{array}{l} y_0 = \eta \\ y_{n+1} = y_n + h\Phi(t_n, y_n; h)(n = 0, 1, ..., N-1) \end{array} \tag{6.1}$$

donde ϕ es una función que depende de la ecuación diferencial considerada y se llama función incremento del método, puede pensarse como el incremento aproximado por unidad de paso. Para el denominado método de **Euler explícito o progresivo**, descrito antes es $\phi(t, y; h) = f(t, y)$ y para el método de **Euler implícito o regresivo**, dado por el algoritmo

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1})$$

la función incremento ϕ habría que definirla implícitamente por medio de la ecuación funcional

$$\phi(t, y; h) = f(t + h, y + h\phi(t, y; h))$$

En general, si la función incremento se puede expresar explícitamente en términos de f y sus derivadas el método se llama explícito, en caso contrario se dice implícito. Sin pérdida de generalidad, tomaremos redes uniformes, con paso fijo $h = t_{j+1} - t_j = \frac{b-a}{N} > 0$ (en otro caso basta con hacer que $h = \max h_n \longrightarrow 0$). Dada la ecuación diferencial $y'(t) = f(t, y(t))$, supondremos, como siempre, que f es continua en D y que verifica una condición de Lipschitz uniforme con respecto a y en D o dicho brevemente que $f \in C_L$.

Definición 21 *Llamaremos* ***error local*** *en el punto* $(\hat{t}, \hat{y})$ *con paso* h *al que se cometería partiendo de ese punto con paso* h*, dividido por* h*, es decir*

$$e(\hat{t}, \hat{y}; h) = \frac{1}{h}[y(\hat{t} + h; \hat{t}, \hat{y}) - \hat{y}] - \Phi((\hat{t}, \hat{y}; h)$$

o sea es la diferencia entre el incremento exacto y aproximado por unidad de paso. En algunos textos se define simplemente como $e(\hat{t}, \hat{y}; h) = y(\hat{t}+h; \hat{t}, \hat{y}) - \hat{y} - h\Phi((\hat{t}, \hat{y}; h)$.

Se dirá que este error es de ***orden*** *p, si este es el mayor entero positivo tal que para toda ecuación diferencial con f de clase $C^{(p)}$, continua y lipschitziana con respecto a y en D, se tiene*

$$| e(\hat{t}, \hat{y}; h) |= O(h^p), (h \to O^+)$$

Si la definición de error local fuese la segunda citada antes en la definición de orden p debería figurar $| e(\hat{t}, \hat{y}; h) |= O(h^{p+1}), (h \to O^+)$.

El concepto de **estabilidad** se introduce para controlar el comportamiento global del método frente a perturbaciones locales, que pueden interpretarse como errores locales de truncatura o de redondeo. Para simplificar la presentación, en lo sucesivo, y mientras no se diga lo contrario, supondremos que el método se aplica con paso fijo $h = T/N$, donde N es un entero positivo.

Definición 22 *Sea el esquema numérico original*

$$\begin{cases} y_0 = \eta \\ y_{n+1} = y_n + h\Phi(t_n, y_n; h)(n = 0, 1, ..., N-1) \end{cases} \tag{6.2}$$

y el esquema perturbado

$$\begin{cases} \hat{y}_0 = y_0 + \omega_0 \\ \hat{y}_{n+1} = \hat{y}_n + h\Phi(t_n, \hat{y}_n; h) + \omega_{n+1}(n = 0, 1, ..., N-1) \end{cases} \tag{6.3}$$

donde ω_i son perturbaciones arbitrarias. Diremos que el método (6.1) es ***estable****, si para todo PVI las soluciones de (6.2) y (6.3) verifican*

$$\max_{0 \leq n \leq N} | \hat{y}_n - y_n | \leq k_1 \sum_{j=0}^{N} | \omega_j |$$

donde k_1 depende sólo del PVI considerado, pero no depende de h ni de las perturbaciones.

Con objeto de averiguar si un método (6.1) tiene esta propiedad, exigible a todo método numérico, enunciamos a continuación un teorema que nos da condiciones suficientes para la estabilidad de un método numérico de un paso.

Teorema 38 ***(Condición suficiente de estabilidad)*** *Si la función incremento del método numérico $\Phi(t, y; h)$ verifica una condición de Lipschitz respecto a y de la forma: existe L constante positiva tal que $\forall y, \hat{y} \in \mathbb{R}^s$, $t \in [a, b]$ y $h \in [0, h_0)$ es $| \Phi(t, y; h) - \Phi(t, \hat{y}; h) | \leq L | y - \hat{y} |$; entonces, el método numérico de un paso (6.1) es estable.*

Desde luego a todo método numérico le vamos a requerir que sea convergente, es decir que las aproximaciones que genere converjan a la solución buscada, lo que definimos a continuación con más precisión.

Definición 23 *El método (6.1) se dice* ***convergente*** *si para todo PVI con $f \in C_L$ se verifica que*

$$\lim_{h \to 0^+} \max_{0 \leq n \leq N} \mid y(t_n) - y_n \mid = 0 \ (\textit{siendo } Nh = T)$$

Es decir, si el límite cuando $h \to 0^+$ del mayor de los módulos de las diferencias entre los valores exactos $y(t_n)$ del PVI y los aproximados obtenidos por el método y_n, en cada punto de la red, es igual a cero. Si se denominan errores de discretización global $EG(t_n) = y(t_n) - y_n$, el método es ***convergente*** *si para todo PVI es*

$$\lim_{h \to 0^+} \max_{0 \leq n \leq N} \mid EG(t_n) \mid = 0 \ (\textit{siendo } Nh = T)$$

Y se dice ***convergente de orden p*** *si este es el mayor entero tal que para todo PVI con $f \in C_L \bigcap C^{(p)}$, se verifica que*

$$\max_{0 \leq n \leq N} \mid y(t_n) - y_n \mid = \max_{0 \leq n \leq N} \mid EG(t_n) \mid = O(h^p) \ (h \to 0^+)$$

Nos será útil introducir un nuevo concepto, a saber la consistencia, que nos mide cómo las soluciones exactas verifican la ecuación del método numérico.

Definición 24 *El método (6.1) se dice* ***consistente*** *si para todo PVI con $f \in C_L$ se verifica que*

$$\lim_{h \to 0^+} \max_{0 \leq n \leq N-1} \mid \frac{1}{h}[y(t_{n+1}) - y(t_n)] - \Phi(t_n, y(t_n); h) \mid = 0 \ (\textit{siendo } Nh = T)$$

para toda solución $y(t)$ de la ecuación diferencial $y' = f(t, y)$. Y ***consistente de orden p*** *si este es el mayor entero tal que para todo PVI con $f \in C_L \bigcap C^{(p)}$, se verifica que*

$$\max_{0 \leq n \leq N-1} \mid \frac{1}{h}[y(t_{n+1}) - y(t_n)] - \Phi(t_n, y(t_n); h) \mid = O(h^p) \ (h \to 0^+)$$

Ahora estamos en condiciones de enunciar el teorema fundamental que sigue, que nos da una condición necesaria y suficiente de convergencia.

Teorema 39 ***(Teorema fundamental).*** *Un método numérico de un paso (6.1), con función incremento verificando la condición de Lipschitz del teorema anterior, es convergente (respectivamente convergente de orden p) si y sólo sí es consistente (respectivamente consistente de orden p).*

Veamos seguidamente el enunciado de dos teoremas que nos dan condiciones suficientes de consistencia.

Teorema 40 *Si el error local de un método (6.1) es de orden p, el método es consistente de orden p, y por el anterior convergente del mismo orden.*

Teorema 41 *Si la función incremento de un método (6.1), Φ, es continua respecto a todos sus argumentos en su dominio de definición, y verifica la condición de Lipschitz del teorema 35. El método (6.1) es consistente (y por lo tanto convergente) si y sólo si para todo PVI con $f \in C_L$ se tiene que*

$$\Phi(t, y, 0) = f(t, y), \ \forall (t, y) \in [a, b] \times \mathbb{R}^s$$

Los conceptos de convergencia, estabilidad y consistencia se generalizan sin dificultad a redes no uniformes. Siendo los teoremas anteriores igualmente válidos en esta situación más general.

Finalmente, para el estudio del orden de un método numérico de un paso, nos será muy útil el teorema siguiente, que se prueba desarrollando en potencias de h, mediante desarrollos de Taylor, los errores locales, con paso h, a partir de los puntos $(t_n, y(t_n))$ de la solución exacta.

Teorema 42 *Sea $f : [a, b] \times \mathbb{R}^s \to \mathbb{R}^s$ de clase $C^{(p)}$ y suponer que las funciones $\Phi(t, y; h)$, $\frac{\partial \Phi}{\partial h}(t, y; h)$, ... , $\frac{\partial^p \Phi}{\partial h^p}(t, y; h)$ son continuas en $[a, b] \times \mathbb{R}^s \times [0, h^*]$, siendo $\Phi(t, y; h)$ lipschitziana con respecto a y en dicho dominio. Entonces, una condición necesaria y suficiente para que el método numérico de función incremento $\Phi(t, y; h)$ sea de orden, al menos, p es que se verifiquen las condiciones siguientes:*

$$\begin{gathered} \Phi(t, y; 0) = f(t, y) \\ \frac{\partial \Phi}{\partial h}(t, y; 0) = \frac{1}{2} f^{(1)}(t, y) \\ \dots \\ \frac{\partial^{p-1} \Phi}{\partial h^{p-1}}(t, y; 0) = \frac{1}{p} f^{(p-1)}(t, y) \end{gathered}$$

Y su orden será exactamente p si además es

$$\frac{\partial^p \Phi}{\partial h^p}(t, y; 0) \neq \frac{1}{p+1} f^{(p)}(t, y)$$

donde las derivadas sucesivas de f a lo largo de las soluciones del sistema diferencial vienen dadas por

$$\begin{gathered} f^{(0)}(t, y) = f(t, y) \\ f^{(1)}(t, y) = f_t^{(0)}(t, y) + f_y^{(0)}(t, y) f(t, y) \\ \dots \\ f^{(k)}(t, y) = f_t^{(k-1)}(t, y) + f_y^{(k-1)}(t, y) f(t, y) \end{gathered}$$

6.3. Métodos de Taylor

El método más sencillo de un paso es el método de Euler, que a su vez es fácilmente programable, pero tiene el inconveniente de que al ser los errores de discretización global (EG) del orden de h, para conseguir que sean pequeños hay que tomar pasos pequeños, con lo que el costo computacional se hace mayor, a la vez que los errores de redondeo aumentan. Interesa pues, buscar métodos que den EG pequeños con un costo computacional razonable. Entonces, se hace necesario considerar métodos de orden $p > 1$. Una posibilidad es la siguiente:

Métodos de Taylor de orden p. Si la función que define la ecuación diferencial es $f \in C_L \bigcap C^{(p)}$; entonces, la solución exacta del PVI

$$PVI \left\{ \begin{array}{l} y'(t) = f(t, y(t)), \forall t \in [a, b] \\ y(t_n) = y_n \end{array} \right.$$

admite el siguiente desarrollo de Taylor si $t_n, t_n + h \in [a, b]$:

$$y(t_n+h; t_n, y_n) = y_n+hy'(t_n)+\frac{h^2}{2!}y''(t_n)+\cdots+\frac{h^p}{p!}y^{(p)}(t_n)+\frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\hat{t}_n) =$$

$$y_n + hf(t_n, y_n) + \frac{h^2}{2!}f^{(1)}(t_n, y_n) + \cdots + \frac{h^p}{p!}f^{(p-1)}(t_n, y_n) + \frac{h^{p+1}}{(p+1)!}f^{(p)}(\hat{t}_n, \hat{y}_n)$$

siendo $\hat{t}_n$ es un punto intermedio entre t_n y t_{n+1}, y donde

$$\begin{array}{c} y'(t_n) = f(t_n, y_n) \\ y''(t_n) = f^{(1)}(t_n, y_n) = (f_t + f_y f)(t_n, y_n) \\ y'''(t_n) = f^{(2)}(t_n, y_n) = (f_t^{(1)} + f_y^{(1)} f)(t_n, y_n) \\ \cdots \\ y^{(k+1)}(t_n) = f^{(k)}(t_n, y_n) = (f_t^{(k-1)} + f_y^{(k-1)} f)(t_n, y_n) \end{array}$$

Por tanto, si prescindimos del resto en la expresión del desarrollo de Taylor anterior, obtenemos el método numérico de un paso

$$y_{n+1} = y_n+h\Phi(t_n, y_n; h) = y_n+hf(t_n, y_n)+\frac{h^2}{2!}f^{(1)}(t_n, y_n)+\cdots+\frac{h^p}{p!}f^{(p-1)}(t_n, y_n)$$

que verifica

$$\mid \frac{1}{h}[y(t_n + h; t_n, y_n) - y_n] - \Phi(t_n, y_n; h) \mid = O(h^p), (h \longrightarrow 0^+)$$

es decir el error local es de orden p, se denomina **método de Taylor de orden** p. Para el caso $p = 1$ se reduce al método de Euler. Desde luego, la función incremento para el método de Taylor de orden p está dada por:

$$\Phi(t, y; h) = f(t, y) + \frac{h}{2!} f^{(1)}(t, y) + \cdots + \frac{h^{p-1}}{p!} f^{(p-1)}(t, y)$$

Es fácil ver que si las funciones $f^{(k)}(t, y)$ verifican en su dominio $[a, b] \times \mathbb{R}^s$, condiciones de Lipschitz con respecto a y, con constates de Lipschitz respectivas L_0, L_1,...;L_{p-1}; entonces, $\Phi(t, y; h)$ también la verifica con constante de Lipschitz:

$$L = L_O + \frac{1}{2} L_1 + \cdots + \frac{1}{p!} L_{p-1}$$

en virtud de los teoremas anteriores se concluye que se trata de un **método convergente de orden** p. Así pues, si la función que describe la ecuación diferencial f, es una función suficientemente derivable, en principio, disponemos de un procedimiento sencillo para obtener métodos numéricos de integración de un paso de orden elevado. Aunque este método puede presentar los siguientes inconvenientes: 1) Que f(t,y) no admita las derivadas necesarias para aplicarlo. 2) Que aún existiendo las derivadas sucesivas, la complejidad de su cálculo sea extraordinariamente grande y esto complique o haga imposible su obtención efectiva.

Ejercicio. Aplicar el método de Taylor de tercer orden para integrar numéricamente el PVI: $y' = y, y(0) = 1$, tomando $h = 0{,}1$ en el intervalo $[0, 1]$.

Hemos de realizar el algoritmo

$$\begin{aligned} &y_0 = 1 \\ &y_{n+1} = y_n + h\Phi(t_n, y_n; h)(n = 0, 1, ..., N-1) \end{aligned} \tag{6.4}$$

donde para orden tres, la función incremento se reduce a

$$\Phi(t_n, y_n; h) = f(t_n, y_n) + \frac{h}{2!} f^{(1)}(t_n, y_n) + \frac{h^2}{3!} f^{(2)}(t_n, y_n)$$

y puesto que en este caso se tiene que

$$f(t, y) = f^{(1)}(t, y) = f^{(2)}(t, y) = y$$

la función incremento viene dada por

$$\Phi(t_n, y_n; h) = y_n + \frac{h}{2} y_n + \frac{h^2}{3!} y_n = (1 + \frac{h}{2} + \frac{h^2}{3!}) y_n$$

y el método se reduce a realizar el algoritmo

$$\begin{aligned} y_0 &= 1 \\ y_{n+1} &= (1 + h + \tfrac{h^2}{2} + \tfrac{h^3}{3!}) y_n (n = 0, 1, ..., 9) \end{aligned} \tag{6.5}$$

ahora sustituyendo cada uno en el siguiente y teniendo en cuenta que $y_0 = 1$, resulta

$$y_{n+1} = (1 + h + \frac{h^2}{2} + \frac{h^3}{3!})^{n+1} (n = 0, 1, ..., 9)$$

Reteniendo 6 cifras decimales redondeadas se obtiene la siguiente tabla, en cuya primera columna se representan los nodos $t_i = i * 0{,}1$, en la segunda los valores aproximados y_i, proporcionados por el método de Taylor de tercer orden, y en la tercera los valores de la solución exacta $y(t_i) = e^{t_i}$

t_i	y_i	$y(t_i)$
0,0	1,000000	1,000000
0,1	1,105167	1,105171
0,2	1,221393	1,221403
0,3	1,349843	1,349859
0,4	1,491802	1,491825
0,5	1,648690	1,648721
0,6	1,822077	1,822119
0,7	2,013698	2,013753
0,8	2,225472	2,225541
0,9	2,459518	2,459603
1,0	2,718177	2,718282

Que aún coincidiendo en $x = 1$ sólo tres cifras decimales de la solución aproximada obtenida y de la exacta mejora notablemente al de Euler que, integrando con el mismo paso, sólo da correcta la parte entera.

6.4. Desarrollo asintótico del error global y aplicaciones

Para el estudio de desarrollos asintóticos del error en los métodos generales de un paso, haremos las siguientes hipótesis sobre el PVI a integrar y el método de integración:

1. Que $f \in C^{(p)}(\Omega)$, siendo Ω el ϵ-entorno de la solución $\Omega = \{(t, y) | t \in [a, b] \wedge |y - y(t)| \leq \epsilon\}$.

2. Que el método numérico de integración tiene orden $p \geq 1$ y su función incremento $\Phi(t, y; h) \in C^{(q)}(\Omega \times [0, h^*])$ y $p + 1 \leq q$.

Entonces podemos enunciar el siguiente teorema.

Teorema 43 *Si el PVI y el método numérico verifican las hipótesis 1 y 2 anteriores, el error global en todo punto $t = t_n$ de la red admite un desarrollo asintótico de la forma:*

$$y(t) - y_h(t) = d_p(t)h^p + \cdots + d_q(t)h^q + E(t, h)h^{q+1}$$

donde las funciones $d_i(t)$ depende de t pero no de h y $E(t, h)$ está acotada en conjuntos compactos.

6.4.1. Estimación del error global mediante el método de extrapolación al límite de Richardson

Sea $y_h(t)$ el valor de la solución numérica en el punto t, calculado con paso h, donde $t = t_n = nh$ es un punto de la red. De acuerdo con el teorema anterior, si el método es de orden p, se tiene:

$$y(t) - y_h(t) = d_p(t)h^p + O(h^{p+1})$$

siendo $d_p(t)$ independiente del paso h, repitiendo la integración con paso $h/2$ se tendrá:

$$y(t) - y_{h/2}(t) = d_p(t)(h/2)^p + O(h^{p+1})$$

y restando a la primera la segunda, resulta

$$y_{h/2}(t) - y_h(t) = (1 - \frac{1}{2^p})h^p d_p(t) + O(h^{p+1}) = (2^p - 1)(h/2)^p d_p(t) + O(h^{p+1})$$

luego

$$\frac{y_{h/2}(t) - y_h(t)}{2^p - 1} = (h/2)^p d_p(t) + O(h^{p+1})$$

por tanto

$$E = \frac{y_{h/2}(t) - y_h(t)}{2^p - 1}$$

es una **estimación asintóticamente correcta** del error de la solución de orden p con paso $h/2$ y se llama de Richardson.

En realidad, seguún el teorema anterior, el error global está dado por un desarrollo de la forma

$$y(t) - y_h(t) = d_p(t)h^p + d_{p+1}(t)h^{p+1} + \ldots$$

donde las $d_i(t)$ son independientes del paso h; por tanto, se tendrá

$$y(t) - y_{h/2}(t) = d_p(t)(h/2)^p + d_{p+1}(t)(h/2)^{p+1} + \dots$$

ahora, multiplicando esta segunda por 2^p y restándole la primera se obtiene

$$(2^p - 1)y(t) - 2^p y_{h/2}(t) + y_h(t) = -(\frac{1}{2})^p d_{p+1}(t)h^{p+1} + \dots$$

por tanto

$$y(t) - \frac{2^p y_{h/2}(t) - y_h(t)}{2^p - 1} = -(\frac{1}{2(2^p - 1})d_{p+1}(t)h^{p+1} + \dots$$

luego

$$AM = \frac{2^p y_{h/2}(t) - y_h(t)}{2^p - 1} = \frac{y_h(t) - 2^p y_{h/2}(t)}{1 - 2^p} = y_{h/2}(t) + E$$

es una aproximación de $y(t)$ de orden superior a las previas (concretamente de orden mayor o igual que $p + 1$), que podemos llamar **aproximación mejorada de Richardson**. Este procedimiento se puede reiterar y obtener aproximaciones de orden superior, siendo aplicable a otros problemas cuyo desarrollo del error es similar al expuesto, lo vemos abreviadamente en lo que sigue.

6.4.2. Extrapolación al límite de Richardson

En muchas situaciones en Análisis Numérico queremos evaluar un número I_0, pero sólo somos capaces de obtener una aproximación $I(h)$, donde h es un parámetro de discretización positiva (longitud de paso) y donde $I(h) \to I_0$ cuando $h \to 0$; supongamos además que $I(h)$ tiene un desarrollo asintótico de la forma

$$I(h) = \sum_{k=0}^{N} I_k h^k + O(h^{N+1})(h \to 0)$$

donde los coeficientes I_k son independientes de h, que escribiremos de la forma

$$I(h) \cong \sum_{k=0}^{N} I_k h^k$$

supongamos que calculamos $I(h_0)$ e $I(h_0/2)$, entonces

$$I(h_0) = I_0 + I_1 h_0 + I_2 h_0^2 + \dots = I_0 + O(h_0)$$

$$I(h_0/2) = I_0 + I_1h_0/2 + I_2(h_0/2)^2 + \cdots = I_0 + O(h_0/2)$$

combinando ambas

$$2I(h_0/2) - I(h_0) = I_0 - \frac{1}{2}I_2h_0^2 + \cdots = I_0 + O(h_0^2)$$

Luego, $2I(h_0/2) - I(h_0)$, es una mejor aproximación de I_0, que $I(h_0)$ o $I(h_0/2)$, Si combinamos $I(h_0)$, $I(h_0/2)$, $I(h_0/4)$, podemos obtener una aproximación de I_0 con un error $O(h_0^3)$. No es necesario usar la sucesión $h_0, h_0/2, h_0/4$, es suficiente con tomar $h_0 > h_1 > h_2 > \ldots$ En general, podemos encontrar una combinación lineal con la propiedad

$$I(h) = \sum_{s=0}^{S} c_{sS}I(h_s) = I_0 + O(h_0^{S+1})(h \to 0)$$

6.5. Métodos Runge-Kutta explícitos de m etapas: formulación general

Se llama **método RK explícito de m etapas** (abreviadamente **RK(m)**) para la resolución numérica del problema de valor inicial

$$PVI \begin{cases} y'(t) = f(t, y(t)), \forall t \in I \\ y(t_0) = y_0 \end{cases}$$

a un algoritmo en el que a partir de la aproximación a la solución (t_n, y_n) se obtiene la (t_{n+1}, y_{n+1}) mediante

$$\begin{array}{l} t_{n+1} = t_n + h \\ y_{n+1} = y_n + h\sum_{j=1}^{m} b_iK_i \end{array} \qquad (6.6)$$

donde

$$\begin{array}{l} K_1 = f(t_n, y_n) \\ K_2 = f(t_n + c_2h, y_n + ha_{21}K_1) \\ K_3 = f(t_n + c_3h, y_n + h(a_{31}K_1 + a_{32}K_2)) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ k_m = f(t_n + c_mh, y_n + h(a_{m1}K_1 + a_{m2}K_2 + a_{m,m-1}K_{m-1})) \end{array}$$

c_i, a_{ij}, b_j, son constantes reales características del método. Siguiendo las notaciones de Butcher es habitual especificar un método **RK(m)** por sus

coeficientes escritos en forma de tabla como sigue

$$\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_m & a_{m1} & a_{m2} & \dots & a_{m,m-1} & \\
\hline
& b_1 & b_2 & \dots & b_{m-1} & b_m
\end{array}$$

o abreviadamente

$$\begin{array}{c|c}
c & A \\
\hline
& b^t
\end{array}$$

donde A es una matriz real de orden $m \times m$ con ceros en la diagonal y por encima de esta, b y c son matrices reales de órdenes $m \times 1$, cuyos elementos aparecen en la tabla anterior, en tanto que b^t es la traspuesta de la matriz columna b. El número de etapas m es igual al número de evaluaciones de la función f que se realizan en cada paso.

En la forma usual de un método explícito, se escribirá

$$y_{n+1} = y_n + h\Phi(t_n, y_n; h)$$

con $\Phi(t_n, y_n; h) = \sum_{j=1}^{m} b_j K_j$ y $K_j = f(t_n + c_j h, y_n + h \sum_{s=1}^{j-1} a_{js} K_s$ los definidos anteriormente, por tanto

$$\Phi(t, y; 0) = \sum_{j=1}^{m} b_j f(t, y) = f(t, y) \sum_{j=1}^{m} b_j$$

luego, en las condiciones del teorema 35 será consistente si y sólo si $\sum_{j=1}^{m} b_j = 1$, condición que supondremos siempre.

6.5.1. Tablas de Butcher: ejemplos diversos

1. El método debido a **Runge** (1895) está dado por la tabla

$$\begin{array}{c|cc}
0 & & \\
1/2 & 1/2 & \\
\hline
& 0 & 1
\end{array}$$

que se corresponde con el algoritmo

$$\begin{cases}
K_1 = f(t_n, y_n) \\
K_2 = f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hK_1) \\
y_{n+1} = y_n + hf(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(t_n, y_n))
\end{cases}$$

y se trata de un método Runge-Kutta explícito de dos etapas y orden dos.

2. El método debido a **Kutta** (1905) está dado por la tabla

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

y se corresponde con el algoritmo

$$\left\{ \begin{array}{l} K_1 = f(t_n, y_n) \\ K_2 = f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hK_1) \\ K_3 = f(t_n + \frac{1}{2}h, y_n + \frac{1}{2}hK_2) \\ K_4 = f(t_n + h, y_n + hK_3) \\ y_{n+1} = y_n + \frac{1}{6}h[K_1 + 2(K_2 + K_3) + K_4] \end{array} \right.$$

que es un método RK explícito de cuatro etapas y orden cuatro, es de los más conocidos y utilizados, le denominamos el método **Runge-Kutta "clásico" de cuarto orden**.

Observación.- Respecto a la aplicación del método RK(m) a sistemas diferenciales no autónomos, notemos que dado el PVI no autónomo

$$\left\{ \begin{array}{l} y'(t) = f(t, y(t)), \forall t \in I \\ y(t_0) = y_0 \in \mathbb{R}^s \end{array} \right.$$

si se introducen los vectores de $\mathbb{R}^{s+1}$: $\bar{y}(t) = (t, y(t))^t$, $\bar{y}_0 = (t_0, y_0)^t$ y $\bar{f}(\bar{y}) = (1, f(t, y(t)))^t$ el problema de valor inicial no autónomo anterior es equivalente al problema autónomo

$$\left\{ \begin{array}{l} \bar{y}'(t) = \bar{f}(\bar{y}(t)), \forall t \in I \\ \bar{y}(t_0) = \bar{y}_0 \end{array} \right.$$

Por tanto, si aplicamos el método RK(m) (6.6) a este, interesa que la solución $\bar{y}_{n+1}(t)$ en t_{n+1} sea de la forma (t_{n+1}, y_{n+1}), donde $t_{n+1} = t_n + h$ e y_{n+1} sean los obtenidos anteriormente; y puede probarse, pero no lo haremos, que esto ocurre para toda ecuación diferencial sí y sólo si los coeficientes del método verifican las relaciones

$$\sum_{j=1}^{m} b_j = 1$$

$$\sum_{j=1}^{i-1} a_{ij} = c_i \ (i = 1, 2, \ldots, m)$$

Por ello, en lo **sucesivo consideraremos únicamente métodos RK(m) verificando** estas igualdades, la primera de las cuales es la **condición de consistencia** y las segundas son denominadas **condiciones de simplificación** (y expresan el hecho de que para cada fila i, el c_i correspondiente debe ser igual a la suma de los a_{ij} de esa misma fila, lo que se traducirá en la simplificación de las condiciones de orden).

6.5.2. Convergencia y orden de los métodos RK(m) explícitos

Consideraremos la clase PVI autónomos de la forma

$$\begin{cases} y'(t) = f(y(t)) \\ y(t_0) = y_0 \end{cases}$$

donde f verifica la condición de Lipschitz con respecto a y, con constante de Lipschitz L y supongamos un método RK(m) dado por

$$y_{n+1} = y_n + h \sum_{j=1}^{m} b_i K_i$$

donde

$$K_i = f(y_n + h \sum_{j=1}^{i-1} a_{ij} K_j) \ (i = 1, 2, \ldots, m)$$

Para aplicar el teorema de estabilidad es necesario asegurar que la función incremento satisface la condición de Lipschitz con respecto a y. Para ello, se puede probar por inducción el siguiente teorema (cuya demostración detallada puede verse en (2)).

Teorema 44 *Si f verifica la condición de Lipschitz con respecto a y, con constante L para todo $h \in (0, h_0)$, entonces $\phi(y, h)$ también la verfica con constante $\bar{L} = \beta_1 L(1 + h_0 La)^{m-1}$, donde $a = \text{máx} \mid a_{ij} \mid$ y $\beta_1 = \sum_{j=1}^{m} \mid \beta_j \mid$.*

Conclusión.- Por el teorema fundamental de convergencia, todo método RK(m) explícito es convergente si y sólo si es consistente (y convergente de orden q si y sólo si es consistente de orden q. Que es consistente es inmediato pues $\phi(t, y; 0) = f(t, y)$ luego todos estos métodos son convergentes.

Orden de los métodos Runge-Kutta explícitos. Barreras de Butcher

El estudio del orden es algo complicado y recurre, en general, digamos que para un número de etapas mayor o igual que 4, a la teoría de árboles. Seguidamente, damos sin demostración algunos resultados relativos al orden máximo alcanzable por un método Runge-Kutta explícito de m etapas.

Teorema 45 *Todo método RK explícito de m etapas tiene orden menor o igual que m. Además, en la tabla que sigue se muestran algunos resultados relativos al orden máximo alcanzable por un método Runge-Kutta explícito de m etapas (barreras de Butcher):*

Número de etapas (m) =	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	...
Máximo orden alcanzable	*1*	*2*	*3*	*4*	*4*	*5*	*6*	*6*	*7*	*7*	*8*	...

Condiciones de orden.- Se pueden obtener, ya sea por desarrollo en serie de Taylor de los errores o por aplicación del teorema 7 sobre el orden, las siguientes **condiciones de orden 2 para los métodos Runge-Kutta explícitos de 2 etapas** verificando la condición de simplificación $c_2 = a_{21}$:

$$\begin{array}{l} b_1 + b_2 = 1 \\ b_2 c_2 = \frac{1}{2} \end{array} \tag{6.7}$$

de donde se obtiene la familia uniparamétrica de métodos RK explícitos de 2 etapas y orden 2

$$b_2 = \frac{1}{2c_2};\ b_1 = 1 - \frac{1}{2c_2}\ (c_2 = a_{21},\ \textit{constante no nula}) \tag{6.8}$$

Y las **condiciones de orden 3 para los métodos Runge-Kutta explícitos de 3 etapas**, verificando las condiciones de simplificación $c_2 = a_{21}$ y $c_3 = a_{31} + a_{32}$, resultan ser:

$$\begin{array}{l} 1)\ b_1 + b_2 + b_3 = 1 \\ 2)\ b_2 c_2 + b_3 c_3 = 1/2 \\ 3)\ b_2 c_2^2 + b_3 c_3^2 = 1/3\ y\ 4)\ \ b_3 a_{32} c_2 = 1/6 \end{array} \tag{6.9}$$

Así, el método es de orden 1 si se verifica la primera, pero no la segunda, de orden 2 si se verifican primera y segunda pero no alguna de las restantes y de orden 3 (máximo alcanzable) si se verifican todas.

6.6. Formulación general de los métodos lineales multipaso: orden, consistencia, estabilidad y convergencia

Seguidamente estudiaremos los métodos lineales multipaso (MLM) para la resolución numérica del PVI:

$$\begin{cases} y'(t) = f(t, y(t)), \forall t \in I \\ y(t_0) = y_0 \end{cases} \tag{6.10}$$

En los métodos multipaso la integración avanza paso a paso utilizando información de la solución en varios pasos previos, concretamente los valores de la solución y/o la derivada en varios puntos consecutivos de la red; así, supongamos calculadas las aproximaciones y_j, $y'_j = f_j = f(t_j, y_j)$ $(j = 0, 1, \ldots, n-1)$ a la solución y su derivada en los puntos t_j de una red en el intervalo $[a, b]$. Un **método de k pasos** es un algoritmo que permite determinar $y_n \simeq y(t_n)$ en función de y_{n-j}, f_{n-j} $(j = 1, \ldots, k)$. Con objeto de simplificar supondremos que la red es uniforme con paso fijo h, de manera que $t_j = a + jh$ $(j = 0, 1, \ldots, N)$ (siendo $Nh = T = b - a$); y nos restringiremos a los métodos lineales multipaso con coeficientes constantes, en los que el algoritmo que define y_n es lineal (se dice lineal porque los valores de y_j, y'_j aparecen linealmente), y adopta la forma

$$\textstyle\sum_{j=0}^{k} \alpha_{k-j} y_{n-j} = h \sum_{j=0}^{k} \beta_{k-j} f_{n-j} \tag{6.11}$$

donde α_j, β_j son coeficientes constantes independientes del paso h, siendo $\alpha_k \neq 0$ y $\mid \alpha_0 \mid + \mid \beta_0 \mid \neq 0$. Si $\beta_k = 0$, la fórmula se dice **explícita**, y en caso contrario, es decir si $\beta_k \neq 0$ se dice **implícita**. La fórmula anterior también puede escribirse desarrollada como sigue

$$\alpha_k y_n + \alpha_{k-1} y_{n-1} + \cdots + \alpha_0 y_{n-k} = h(\beta_k f_n + \beta_{k-1} f_{n-1} + \cdots + \beta_0 f_{n-k}) \tag{6.12}$$

Del problema de valor inicial sólo se conoce el valor de la solución en el punto inicial, por tanto un método de k pasos $(k > 1)$ no puede ser aplicado directamente y es necesario dar otro algoritmo (por ejemplo, basado en un método de un paso) para calcular y_j, $y'_j = f_j$, aproximaciones de la solución y su derivada en t_j $(j = 1, \ldots, k-1)$, dicho algoritmo se llama de **iniciación o arranque** y estará implícito en todo método de k pasos. Si el método es explícito, el cálculo desde (6.11) o (6.12) es inmediato a partir de los valores en los k puntos anteriores. Pero si el método fuera implícito se tendría

$$\alpha_k y_n + \alpha_{k-1} y_{n-1} + \cdots + \alpha_0 y_{n-k} = h(\beta_k f_n + \beta_{k-1} f_{n-1} + \cdots + \beta_0 f_{n-k})$$

de donde se deduce que

$$y_n = [h\beta_k f(t_n, y_n) - \sum_{s=1}^{k} \alpha_{k-s} y_{n-s} + h \sum_{s=1}^{k} \beta_{k-s} f(t_{n-s}, y_{n-s})]/\alpha_k$$

luego y_n será solución de la ecuación (generalmente no lineal) en y

$$y = g(y) = h\frac{\beta_k}{\alpha_k} f(t_n, y) + F_n \tag{6.13}$$

donde F_n es una función constante que depende de términos conocidos de los pasos anteriores y está dada por la expresión

$$F_n = [-\sum_{s=1}^{k} \alpha_{k-s} y_{n-s} + h \sum_{s=1}^{k} \beta_{k-s} f(t_{n-s}, y_{n-s})]/\alpha_k$$

Puesto que f es lipschitziana respecto a y, con constante de Lipschitz $L > 0$, será

$$\mid g(y) - g(\bar{y}) \mid \leq h \mid \frac{\beta_k}{\alpha_k} \mid L \mid y - \bar{y} \mid$$

por tanto si h es tal que

$$h \mid \frac{\beta_k}{\alpha_k} \mid L < 1 \Leftrightarrow h < \mid \frac{\alpha_k}{\beta_k} \mid \frac{1}{L}$$

entonces la función $g(y)$ es contractiva y la ecuación (6.13) tiene solución única, que se obtiene como el límite de la sucesión dada por el algoritmo

$$\begin{cases} y_n^0 \ (valor\ inicial) \\ y_n^{k+1} = g(y_n^k) = h\frac{\beta_k}{\alpha_k} f(t_n, y_n^k) + F_n \end{cases}$$

en ocasiones si se parte de un valor inicial adecuado es suficiente con una única iteración, a veces se toma como valor inicial el proporcionado por un método explícito, si el valor de h anterior fuera muy pequeño se podría optar por otro método iterativo, por ejemplo el método de Newton, se introducen en este contexto los denominados método predictor-corrector.

6.6.1. Métodos predictor-corrector

Son pares de métodos lineales multipaso, uno explícito y otro implícito, normalmente del mismo orden, donde la fórmula explícita es utilizada para predecir la siguiente aproximación y la implícita para corregirla. Suponer que utilizamos un método de k pasos, de orden k, explícito con coeficientes α_k, β_k, para el predictor, y un método implícito de $k-1$ pasos para el corrector

con coeficientes α_k^*, β_k^* entonces, si $y_n(0)$ es la aproximación predicha, uno procede como sigue (en el supuesto de que $\alpha_k = \alpha_k^* = 1$):

$$\begin{cases} y_n^0 = -\sum_{s=1}^k \alpha_{k-s} y_{n-s} + h \sum_{s=1}^k \beta_{k-s} f_{n-s} \\ y_n = -\sum_{s=1}^k \alpha_{k-s}^* y_{n-s} + h[\beta_{k-s}^* f(t_n, y_n^0) + \sum_{s=1}^{k-1} \beta_{k-s}^* f_{n-s}] \\ f_n = f(t_n, y_n) \end{cases} \tag{6.14}$$

Esto requiere dos evaluaciones de f por paso y es frecuentemente referido como un PECE method (Predecir, Evaluar, Corregir, Evaluar). Uno puede, naturalmente corregir una vez más y luego salir o reevaluar f, y así sucesivamente. Así pues, hay métodos P(EC)2, P(EC)2E, los más económicos son los del tipo PECE.

6.6.2. Orden, consistencia, estabilidad y convergencia de un método lineal multipaso

Denotando por E el operador desplazamiento, cuyas potencias actúan sobre sucesiones en la forma: $E^j(z_n) = z_{n+j}$, e introduciendo los polinomios característicos del método

$$\begin{array}{l} \rho(\lambda) = \sum_{j=0}^k \alpha_j \lambda^j = \alpha_0 + \alpha_1 \lambda + \alpha_2 \lambda^2 + \cdots + \alpha_k \lambda^k \\ \sigma(\lambda) = \sum_{j=0}^k \beta_j \lambda^j = \beta_0 + \beta_1 \lambda + \beta_2 \lambda^2 + \cdots + \beta_k \lambda^k \end{array} \tag{6.15}$$

y los operadores polinomiales

$$\rho(E) = \sum_{j=0}^k \alpha_j E^j, \quad \sigma(E) = \sum_{j=0}^k \beta_j E^j \tag{6.16}$$

Las ecuaciones del método (6.10) pueden escribirse en la forma abreviada:

$$\rho(E) y_{n-k} = h\sigma(E) f_{n-k} \ (n = k, \ldots, N) \tag{6.17}$$

y se habla del método (ρ, σ).

Definición 25 *Si $y(t)$ es la solución exacta del problema de valor inicial, se define el* ***error de discretización local*** *en t_n de un método lineal multipaso, y se pone* ***EDL***(t_n), *mediante la fórmula*

$$\sum_{j=0}^k \alpha_{k-j} y(t_{n-j}) = h \sum_{j=0}^k \beta_{k-j} y'(t_{n-j}) + \alpha_k EDL(t_n) \tag{6.18}$$

es decir, es el defecto de la solución exacta respecto a la ecuación del método. En relación con el ***EDL***, *para funciones y(t) suficientemente derivables en*

$[a, b]$, *se introduce un operador diferencial lineal* $\boldsymbol{L} = \boldsymbol{L[y(t);h]}$, *que llamaremos* ***operador error local*** *por la expresión*

$$L[y(t); h] = [\textstyle\sum_{j=0}^{k} \alpha_{k-j} y(t - jh) - h\beta_{k-j} y'(t - jh)] \qquad (6.19)$$

Por tanto, se tiene que $EDL(t_n) = L[y(t_n); h]/\alpha_k$. *Además, el operador lineal y el método lineal multipaso* (ρ, σ), *se dicen de* ***orden*** p *si este es el máximo entero positivo tal que*

$$L[y(t); h] = O(h^{p+1}) \text{ (cuando } h \longrightarrow 0^+) \ \forall y(t) \in C^{p+1}([a, b])$$

o equivalentemente

$$\frac{1}{h} L[y(t); h] = O(h^p) \text{ (cuando } h \longrightarrow 0^+) \ \forall y(t) \in C^{p+1}([a, b])$$

El siguiente teorema da condiciones necesarias y suficientes de orden para un método lineal multipaso.

Teorema 46 *Un método lineal multipaso* (ρ, σ) *es de orden* p *sí y sólo si*

$$c_0 = \sum_{j=0}^{k} \alpha_j = 0, \ c_1 = \sum_{j=0}^{k} (j\alpha_j - \beta_j) = 0, \ c_2 = \frac{1}{2!} \sum_{j=0}^{k} (j^2 \alpha_j - 2j\beta_j) = 0$$

$$\ldots, \ c_p = \frac{1}{p!} \sum_{j=0}^{k} (j^p \alpha_j - pj^{p-1} \beta_j) = 0 \text{ y } c_{p+1} \neq 0$$

la c_{p+1} *es denominada constante de error. Si* $c_0 = c_1 = \cdots = c_p = 0$ *pero no precisamos el valor de* c_{p+1}, *decimos que el método tiene orden al menos* p.

Definición 26 *Un método se dice* ***consistente*** *si para todo* $f \in C_L$ *y toda solución* $y(t)$ *de la ecuación diferencial* $y'(t) = f(t, y(t))$, *se verifica*

$$\text{lím}_{h \to 0^+} \text{máx}_{k \leq n \leq N} \mid \textstyle\sum_{j=0}^{k} [\frac{1}{h} \alpha_{k-j} y(t_{n-j}) - \beta_{k-j} f(t_{n-j}, y(t_{n-j}))] \qquad (6.20)$$

Además, se verifica el siguiente teorema.

Teorema 47 *Un MLM es* ***consistente*** *si y sólo si su orden es* ≥ 1, *es decir si y sólo si* $c_0 = c_1 = 0 \Leftrightarrow \rho(1) = 0$ *y* $\rho'(1) = \sigma(1)$.

La definición de estabilidad es análoga a la dada para métodos de un paso y se caracteriza por el teorema siguiente, también conocido como **condición de Dahlquist** de estabilidad.

Teorema 48 *Un método lineal multipaso es estable si y sólo verifica la* ***condición de las raíces****, es decir si toda raíz* λ *de la ecuación* $\rho(\lambda) = 0$ *verifica que* $\mid \lambda \mid < 1$ *o si* $\mid \lambda \mid = 1$ *dicha raíz es simple.*

Definición 27 *Un método se dice* ***convergente*** *si para todo* $f \in C_L$ *y toda solución* $y(t)$ *de la ecuación diferencial* $y'(t) = f(t, y(t))$*, se verifica*

$$\lim_{h\to 0^+} \max_{k\leq n\leq N=[\frac{T}{h}]} \mid y(t_n) - y_n \mid = 0 \qquad (6.21)$$

siempre que los valores de inicialización η_n *satisfagan*

$$\lim_{h\to 0^+} \mid y(t_n) - \eta_n \mid = 0 \ (n = 0, 1, \ldots, k-1)$$

Se llaman **errores de discretización global** a las cantidades $EG(t_n) = y(t_n) - y_n$ y el método se dice **de orden p** si

$$\max_{k\leq n\leq N=[\frac{T}{h}]} \mid EG(t_n) \mid = O(h^p)$$

para todo PVI con $f \in C^{(p)}$.

Ahora estamos en condiciones de enunciar los dos teoremas fundamentales que siguen.

Teorema 49 ***(Teorema de convergencia)*** *Un método lineal multipaso es convergente si y sólo si es consistente y estable.*

Teorema 50 ***(Primera barrera de Dahlquist)*** *El orden máximo alcanzable por un método lineal de* k *pasos que sea consistente y estable (i.e. convergente) es* $k+1$ *si* k *es impar y* $k+2$ *si* k *es par. Además, si* $\beta_k \leq 0$ *(en particular si es explícito) el orden máximo alcanzable es* k*.*

Nota: Los métodos de orden $k+2$ son llamados optimales, aunque en la práctica tampoco funcionan bien, por tanto el mejor orden que se puede alcanzar por un método útil de k pasos es $k+1$.

6.7. Fórmulas de Adams

Para introducir las fórmulas de Adams, que son de los métodos lineales multipaso más utilizados, será útil recordar la fórmula de Newton en diferencias regresivas del polinomio de interpolación: sean $t_{\nu-j} = t_\nu - jh$ ($j = 0, 1, \ldots, q$), $q+1$ puntos distintos igualmente espaciados, donde t_ν es el punto de referencia, h es una constante positiva y $f_j = f(t_j)$ son los valores que

toma una función f en esos puntos. Para expresar el polinomio de interpolación de f en diferencias regresivas, conviene introducir la variable $\tau = \frac{t-t_\nu}{h}$, lo que equivale a tomar t_ν como origen y h como unidad de longitud, entonces el polinomio de interpolación $\pi(t)$, de grado a lo más q, que interpola a f en dichos puntos, según vimos, puede escribirse mediante la fórmula de Newton regresiva como sigue:

$$\pi(t) = \sum_{i=0}^{q}(-1)^j \begin{pmatrix} -\tau \\ j \end{pmatrix} \nabla^j f_\nu \tag{6.22}$$

siendo el número combinatorio

$$\begin{pmatrix} -\tau \\ j \end{pmatrix} = \frac{(-\tau)(-\tau-1)\ldots(-\tau-k+1)}{j!} = \frac{(-1)^j\tau(\tau+1)\ldots(\tau+k-1)}{j!}$$

6.7.1. Fórmulas explícitas o de Adams-Bashforth

En primer lugar, introducimos la familia de métodos lineales multipaso de Adams-Bashforth, que como veremos son métodos lineales explícitos de k pasos y orden k. Para ello, supongamos conocidos

$$y_{n-j}, y'_{n-j} = f(t_{n-j}, y_{n-j}) \; (j = 1, 2, \ldots, k)$$

aproximaciones de la solución y su derivada en k puntos consecutivos de la red anteriores a $t_n = t_0 + nh$, puesto que la solución $y(t)$ del PVI (6.10) verifica la ecuación integral

$$y(t_n) = y(t_{n-1}) + \int_{t_{n-1}}^{t_n} y'(t)dt \tag{6.23}$$

podemos sustituir $y(t_{n-1})$ por y_{n-1} y la función subintegral por su polinomio de interpolación $\pi(t)$ en los k valores previos $\{t_{n-1}, t_{n-2}, \ldots, t_{n-k}\}$, que de acuerdo con (6.22), está dado por

$$\pi(t) = \sum_{j=0}^{k-1}(-1)^j \begin{pmatrix} -\tau \\ j \end{pmatrix} \nabla^j y'_{n-1}$$

con $\tau = \frac{t-t_{n-1}}{h}$; por tanto, aproximando la integral de $y'(t)$ por la de su polinomio de interpolación, tendremos

$$\int_{t_{n-1}}^{t_n} y'(t)dt \simeq \int_{t_{n-1}}^{t_n} \pi(t)dt = h\int_0^1 \; \sum_{j=0}^{k-1}(-1)^j \begin{pmatrix} -\tau \\ j \end{pmatrix} \nabla^j y'_{n-1} \; d\tau =$$

$$= h\sum_{j=0}^{k-1} \gamma_j \nabla^j y'_{n-1} \; con \; \gamma_j = (-1)^j \int_0^1 \begin{pmatrix} -\tau \\ j \end{pmatrix} d\tau$$

valores constantes, característicos del método.

Luego, tras las aproximaciones anteriores, el cálculo de y_n, y'_n se realiza por las fórmulas

$$\left\{ \begin{array}{l} y_n = y_{n-1} + h \sum_{j=0}^{k-1} \gamma_j \, \nabla^j \, y'_{n-1} \\ y'_n = f(t_n, y_n) \end{array} \right. \tag{6.24}$$

que se conocen como fórmulas de **Adams-Bashforth** de k pasos, abreviadamente **A-B** de k pasos o simplemente **AB(k)**. Expresando el segundo miembro de (6.24) en función de los valores $\{y'_{n-j} | \; j = 1, 2, \ldots, k\}$, resulta la siguiente expresión de las fórmulas de Adams-Bashforth:

$$\left\{ \begin{array}{l} y_n = y_{n-1} + h \sum_{j=0}^{k-1} \beta_{k,j} y'_{n-1-j} \\ y'_n = f(t_n, y_n) \end{array} \right. \tag{6.25}$$

donde los coeficientes $\beta_{k,j}$ $(j = 0, 1, \ldots, k-1)$ están dados por la fórmula

$$\beta_{k,j} = (-1)^j \left[\left(\begin{array}{c} j \\ j \end{array} \right) \gamma_j + \left(\begin{array}{c} j+1 \\ j \end{array} \right) \gamma_{j+1} + \cdots + \left(\begin{array}{c} k-1 \\ j \end{array} \right) \gamma_{k-1} \right] \tag{6.26}$$

Según se desprende de (6.25), las fórmulas AB(k) son lineales y explícitas y su aplicación sólo necesita una evaluación de función por paso; fueron obtenidas por Adams y Bashforth en 1883.

Función generatriz de los coeficientes del método. Veamos como obtener de forma sencilla los coeficientes del método γ_m, en vez de realizando las integrales anteriores, que son sencillas para los primeros valores, pero cuyo cálculo se hace más largo y tedioso para valores grandes de m. Sea $G(t) = \sum_0^\infty \gamma_m t^m$ la función real analítica que tiene γ_m como coeficientes de su desarrollo de MacLaurin. Entonces, teniendo en cuenta la expresión anterior de estos coeficientes, resulta

$$\begin{array}{l} G(t) = \sum_{m=0}^\infty (-1)^m t^m \int_0^1 \left(\begin{array}{c} -\tau \\ m \end{array} \right) d\tau = \int_0^1 \sum_{m=0}^\infty (-1)^m t^m \left(\begin{array}{c} -\tau \\ m \end{array} \right) d\tau = \\ = \int_0^1 (1-t)^{-\tau} d\tau = -\frac{t}{(1-t) ln(1-t)} \end{array} \tag{6.27}$$

de esta deducimos que

$$-\frac{ln(1-t)}{t} G(t) = \frac{1}{1-t}$$

y de los conocidos desarrollos en serie de MacLaurin de la serie geométrica

$$\frac{1}{1-t} = 1 + t + t^2 + \ldots$$

$$-\frac{ln(1-t)}{t} = 1 + \frac{t}{2} + \frac{t^2}{3} + \dots$$

obtenemos la identidad formal entre las series de polinomios anteriores

$$(1 + \frac{t}{2} + \frac{t^2}{3} + \frac{t^3}{4} + \dots)(\gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3 + \dots) = (1 + t + t^2 + t^3 + \dots)$$

ahora, comparando los coeficientes de las potencias respectivas de t en ambos miembros resulta

$$\gamma_m + \frac{\gamma_{m-1}}{2} + \frac{\gamma_{m-2}}{3} + \dots + \frac{\gamma_0}{m+1} = 1 \ (m = 0, 1, 2, \dots) \tag{6.28}$$

de la que, fácilmente, obtenemos para los primeros valores de m:

m	0	1	2	3	4	5	6	...
γ_m	1	1/2	5/12	3/8	251/720	95/288	19087/60480	...

en consecuencia, los coeficientes (6.26) resultan ser

k	$\beta_{k,0}$	$\beta_{k,1}$	$\beta_{k,2}$	$\beta_{k,3}$	$\beta_{k,4}$
1	1				
2	3/2	$-1/2$			
3	23/12	$-16/12$	5/12		
4	55/24	$-59/24$	37/24	$-9/24$	
5	1901/720	$-2774/720$	2616/720	$-1724/720$	251/720

Error de discretización local. Recordemos que de acuerdo con (6.18), se llama así al defecto de la solución y(t) respecto a la ecuación en diferencias del método. Notemos que, en este caso, se puede interpretar como la diferencia $y(t_n) - y_n$, supuesto que y_n ha sido calculado a partir de los valores exactos de la solución, es decir $y_{n-j} = y(t_{n-j})$ $(j = 1, 2, \dots, k)$, por tanto sería el error cometido en un paso del algoritmo suponiendo que se partiese de valores exactos. Seguidamente, obtenemos una expresión del mismo suponiendo que $y(t)$ es suficientemente diferenciable. Llamando $\pi(t)$ al polinomio de interpolación de $y'(t)$ en los k puntos $\{t_{n-j} | \ j = 1, 2, \dots, k\}$, se tiene (por definición de estos métodos):

$$EDL(t_n) = \int_{t_{n-1}}^{t_n} (y'(t) - \pi(t))dt$$

Y utilizando la fórmula del error de interpolación de Lagrange de $y'(t)$, se deduce que para algún $\xi' \in [t_{n-k}, t_n]$ se tiene

$$EDL(t_n) = h^{k+1}\gamma_k y^{(k+1)}(\xi')$$

Por tanto, salvo errores de redondeo, la integración numérica con **A-B** de orden k permitiría obtener valores exactos para todo PVI cuya solución $y(t)$ fuera un polinomio de grado $\leq k$. El error introducido en un paso del algoritmo es una $O(h^{k+1})$.

6.7.2. Fórmulas implícitas o de Adams-Moulton

De un modo similar al desarrollado en el párrafo anterior, se introduce la familia de métodos lineales implícitos de k pasos y orden $k+1$, conocidas como fórmulas de Adams-Moulton. Pero, ahora, vamos a sustituir la función subintegral de (6.23) por su polinomio de interpolación $\pi^*(t)$ en los $k+1$ puntos $\{t_{n-j}|\ j = 0, 1, 2, \ldots, k\}$, que está dado por la fórmula

$$\pi^*(t) = \textstyle\sum_{j=0}^{k} (-1)^j \begin{pmatrix} -\tau \\ j \end{pmatrix} \nabla^j\, y'_n$$

siendo ahora $\tau = \frac{t-t_n}{h}$. por tanto, aproximando la integral de $y'(t)$ por la de su polinomio de interpolación, tendremos ahora

$$\int_{t_{n-1}}^{t_n} y'(t)dt \simeq \int_{t_{n-1}}^{t_n} \pi^*(t)dt = h\int_{-1}^{0}\ \textstyle\sum_{j=0}^{k} (-1)^j \begin{pmatrix} -\tau \\ j \end{pmatrix} \nabla^j\, y'_n\ d\tau =$$

$$= h\sum_{j=0}^{k} \gamma_j^*\ \nabla^j\, y'_n\ siendo\ \gamma_j^* = (-1)^j \int_{-1}^{0} \begin{pmatrix} -\tau \\ j \end{pmatrix} d\tau$$

valores constantes, característicos del método. Tras las aproximaciones anteriores se obtiene la fórmula

$$\left\{ \begin{array}{l} y_n = y_{n-1} + h\sum_{j=0}^{k} \gamma_j^*\ \nabla^j\, y'_n \\ y'_n = f(t_n, y_n) \end{array} \right. \qquad (6.29)$$

que se conoce como fórmula de **Adams-Moulton** de k pasos, abreviadamente **A-M** de k pasos o simplemente **AM(k)**. Igual que antes, se puede expresar el segundo miembro de (6.29) en función de los valores $\{y'_{n-j}|\ j = 1, 2, \ldots, k\}$, obteniéndose

$$\left\{ \begin{array}{l} y_n = y_{n-1} + h\sum_{j=0}^{k} \beta_{k,j}^* y'_{n-j} \\ y'_n = f(t_n, y_n) \end{array} \right. \qquad (6.30)$$

donde los coeficientes $\beta_{k,j}^*$ $(j = 0, 1, \ldots, k)$ están dados ahora por la fórmula

$$\beta_{k,j}^* = (-1)^j \left[\begin{pmatrix} j \\ j \end{pmatrix} \gamma_j^* + \begin{pmatrix} j+1 \\ j \end{pmatrix} \gamma_{j+1}^* + \cdots + \begin{pmatrix} k \\ j \end{pmatrix} \gamma_k^* \right] \qquad (6.31)$$

Puesto que se trata de un método implícito, se puede proceder como se indicó en la fórmula (6.13) al comienzo del párrafo 6. Asimismo, pueden obtenerse los coeficientes γ_j^* mediante una función generatriz de modo similar al visto para el caso de los métodos AB(k). Si la solución buscada $y(t)$ es suficientemente diferenciable, puede obtenerse para el error de discretización local la expresión

$$EDL(t_n) = h^{k+2}\gamma_{k+1}^* y^{(k+2)}(\xi)$$

para algún $\xi \in [t_{n-k}, t_n]$. Por tanto, se puede decir que la fórmula A-M de k pasos integraría exactamente PVI cuya solución $y(t)$ fuera un polinomio de grado $\leq k+1$.

6.8. Problemas resueltos

1. Dar el algoritmo del método Runge-Kutta explícito de 2 etapas cuyo tablero de Butcher tiene la forma

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

¿cuál es su orden?. Aproximar, mediante dicho método, la solución en el tiempo $t = 0{,}2$ del PVI: $y'(t) = ty(t)$, $y(0) = 1$, dando un único paso de amplitud $h = 0{,}2$ y dos pasos de amplitud $h = 0{,}1$, estimar el error de esta última por el método de extrapolación al límite de Richardson y usarlo para dar una aproximación mejorada de $y(0{,}2)$.

Solución. Al método Runge-Kutta explícito de dos etapas dado por la tabla del enunciado corresponde el algoritmo

$$\begin{aligned} K_1 &= f(t_n, y_n) \\ K_2 &= f(t_n + h, y_n + hK_1) \\ y_{n+1} &= y_n + \tfrac{h}{2}(K_1 + K_2) \end{aligned}$$

Puesto que se verifica la condición de simplificación $c_2 = a_{21}$ y las ecuaciones de orden

$$\begin{aligned} b_1 + b_2 &= 1/2 + 1/2 = 1 \\ b_2 c_2 &= \tfrac{1}{2} \cdot 1 = \tfrac{1}{2} \end{aligned}$$

el método es de orden dos. Para aproximar la solución de la ecuación propuesta, realicemos un único paso de amplitud $h = 0{,}2$; como $t_0 = 0$,

$y_0 = 1$ y $f(t, y) = ty$, entonces se tendrá

$$K_1 = f(t_0, y_0) = f(0, 1) = 0 \cdot 1 = 0$$
$$K_2 = f(t_0 + h, y_0 + hK_1) = f(0{,}2, 1 + 0{,}2 \cdot 0) = f(0{,}2, 1) = 0{,}2 \cdot 1 = 0{,}2$$
$$y_1 = y_0 + \frac{h}{2}(K_1 + K_2) = 1 + \frac{0{,}2}{2}(0 + 0{,}2) = 1 + 0{,}1 \cdot 0{,}2 = \boxed{1{,}02}$$

luego 1,02 es la aproximación de $y(0{,}2)$ calculada por el método dado con paso $h = 0{,}2$, se indica por $y_{0{,}2}(0{,}2)$.

Calculemos ahora dicha aproximación mediante dos pasos de amplitud $h = 0{,}1$, partiendo del mismo punto $(t_0, y_0) = (0, 1)$, para la primera obtendremos

$$K_1 = f(t_0, y_0) = f(0, 1) = 0 \cdot 1 = 0$$
$$K_2 = f(t_0 + h, y_0 + hK_1) = f(0{,}1, 1 + 0{,}1 \cdot 0) = f(0{,}1, 1) = 0{,}1 \cdot 1 = 0{,}1$$
$$y_1 = y_0 + \frac{h}{2}(K_1 + K_2) = 1 + \frac{0{,}1}{2}(0 + 0{,}1) = 1 + 0{,}05 \cdot 0{,}1 = \boxed{1{,}005}$$

en tanto que para la segunda, partiendo de $(t_1, y_1) = (0{,}1, 1{,}005)$ con paso $h = 0{,}1$, tendremos

$$K_1 = f(t_1, y_1) = f(0{,}1, 1{,}005) = 0{,}1005$$
$$K_2 = f(t_1 + h, y_1 + hK_1) = f(0{,}2, 1{,}01505) = 0{,}20301$$
$$y_2 = y_1 + \frac{h}{2}(K_1 + K_2) = 1{,}005 + \frac{0{,}1}{2}(0{,}1005 + 0{,}20301) = \boxed{1{,}0201755}$$

y $1,0201755$ es la aproximación de $y(0{,}2)$ obtenida con paso $h = 0{,}1$, que se indica por $y_{0{,}1}(0{,}2)$. Ahora se puede estimar el error global de esta última aproximación por el método de extrapolación al límite de Richardson, mediante la fórmula

$$y(t) - y_{\frac{h}{2}}(t) \cong \frac{y_{\frac{h}{2}}(t) - y_h(t)}{2^p - 1} = E$$

donde $y(t)$ es el valor de la solución exacta en el punto t, $y_h(t)$ es la aproximación numérica de la solución en dicho punto, obtenida con el método numérico dado de orden p con el paso h, en este caso el orden del método es $p = 2$, e $y_{\frac{h}{2}}(t)$ la aproximación numérica de la solución en el mismo punto, obtenida por el mismo método, pero con paso mitad $\frac{h}{2}$, resultando en el caso que nos ocupa

$$E = \frac{1{,}0201755 - 1{,}02}{2^2 - 1} \cong \boxed{5{,}85 \cdot 10^{-5}}$$

La aproximación mejorada de $y(t)$, por el método de extrapolación al límite de Richardson se obtiene sumando a $y_{\frac{h}{2}}(t)$ la estimación del error E, en este caso resulta ser

$$AM = y_{\frac{h}{2}}(t) + E \cong 1{,}0201755 + 5{,}85 \cdot 10^{-5} = \boxed{1{,}020234}$$

(Con objeto de que podamos comparar las aproximaciones y estimaciones obtenidas, consignemos la solución exacta que viene dada por la función $y = e^{\frac{t^2}{2}}$, cuyo valor en $t = 0{,}2$ es $y(0{,}2) = 1{,}02020134...$).

2. Transforma el problema $y'' = 2t + 4y + y'$, $y(0) = 6$, $y'(0) = 5$, en un sistema de primer orden. Aproxima la solución en el tiempo $t = 0{,}4$ tomando $h = 0{,}2$ e integrándolo mediante el método Runge-Kutta explícito de 2 etapas y orden 2, cuyo tablero de Butcher tiene la forma

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & b_1 & b_2 \end{array}$$

Solución. Transformemos este PVI para una ecuación de segundo orden en un PVI para dos ecuaciones de primer orden, para ello llamaremos $y_1 = y$, $y_2 = y'$ con lo cual el problema dado es equivalente al siguiente:

$$\begin{cases} y_1' = y_2 \\ y_2' = 4y_1 + y_2 + 2t \\ y_1(0) = 6 \\ y_2(0) = 5 \end{cases}$$

o en forma vectorial, llamando $Y = (y_1, y_2)$, $F(t, Y) = (y_2, 4y_1 + y_2 + 2t)$ e $Y_0 = (6, 5)$, se puede escribir en la forma

$$\boxed{Y' = F(t, Y)\ ,\ Y(0) = (6, 5)}$$

de las ecuaciones de orden

$$\begin{array}{l} b_1 + b_2 = 1 \\ b_2 c_2 = b_2 \cdot 1 = \frac{1}{2} \end{array}$$

se deduce que $b_1 = b_2 = \frac{1}{2}$. Ahora realizaremos dos pasos con este método, de amplitud $h = 0{,}2$, partiendo del punto $t_0 = 0$ e $Y_0 = (6, 5)$, el primer paso será

$$\begin{array}{l} K_1 = F(t_0, Y_0) = F(0, (6, 5)) = (5, 4 \cdot 6 + 5 + 2 \cdot 0) = (5, 29) \\ K_2 = F(t_0 + h, Y_0 + hK_1) = F(0{,}2, (6, 5) + 0{,}2 \cdot (5, 29)) = \\ F(0{,}2, (7, 10{,}8)) = (10{,}8, 4 \cdot 7 + 10{,}8 + 2 \cdot 0{,}2) = (10{,}8, 39{,}2) \\ Y_1 = Y_0 + \frac{h}{2}(K_1 + K_2) = (6, 5) + \frac{0{,}2}{2}((5, 29) + (10{,}8, 39{,}2)) \\ = (6, 5) + (1{,}58, 6{,}82) = (7{,}58, 11{,}82) \end{array}$$

en tanto que para el segundo paso, partiendo de $(t_1, Y_1) = (0{,}2, (7{,}58, 11{,}82))$ con paso $h = 0{,}2$, tendremos

$$\begin{aligned} K_1 &= F(t_1, Y_1) = F(0{,}2, (7{,}58, 11{,}82)) = (11{,}82, 42{,}54) \\ K_2 &= F(t_1 + h, Y_1 + hK_1) = (20{,}328, 73{,}784) \\ Y_2 &= Y_1 + \tfrac{h}{2}(K_1 + K_2) = \boxed{(10{,}7948, 23{,}4524)} \end{aligned}$$

3. Dar el algoritmo del método Runge-Kutta explícito de tres etapas cuyo tablero de Butcher tiene la forma

$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array}$$

¿cuál es su orden?. Aproximar, mediante dicho método, la solución en el tiempo $t = 0{,}2$ del PVI: $y'(t) = ty(t)$, $y(0) = 1$, tomando $h = 0{,}1$.

Solución. Al método Runge-Kutta explícito de tres etapas dado por la tabla del enunciado corresponde el algoritmo

$$\begin{aligned} K_1 &= f(t_n, y_n) \\ K_2 &= f(t_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}hK_1) \\ K_3 &= f(t_n + \tfrac{2}{3}h, y_n + \tfrac{2}{3}hK_2) \\ y_{n+1} &= y_n + \tfrac{h}{4}(K_1 + 3K_3) \end{aligned}$$

Puesto que se cumplen las condiciones de simplificación, veamos las ecuaciones de orden dadas por las fórmulas:

$$\begin{aligned} &1)\ b_1 + b_2 + b_3 = \tfrac{1}{4} + 0 + \tfrac{3}{4} = 1 \\ &2)\ b_2c_2 + b_3c_3 = 0 \cdot \tfrac{1}{3} + \tfrac{3}{4}\tfrac{2}{3} = \tfrac{1}{2} \\ &3)\ b_2c_2^2 + b_3c_3^2 = 0 \cdot (\tfrac{1}{3})^2 + \tfrac{3}{4}(\tfrac{2}{3})^2 = 1/3 \\ &4)\ b_3a_{32}c_2 = \tfrac{3}{4}\tfrac{2}{3}\tfrac{1}{3} = \tfrac{1}{6} \end{aligned}$$

como se verifican todas el método es de orden tres, el máximo alcanzable por un Runge-Kutta explícito de tres etapas.

Como nos piden, aproximaremos la solución del problema de valor inicial dado en el punto $t = 0{,}2$ mediante este método con paso $h = 0{,}1$, hemos de realizar dos pasos partiendo de $t_0 = 0$ e $y_0 = 1$, siendo $f(t, y) = ty$. Reteniendo tan sólo seis cifras decimales significativas, para el primer paso se tiene:

$$\begin{aligned} K_1 &= f(t_0, y_0) = 0 \cdot 1 = 0 \\ K_2 &= f(t_0 + \tfrac{0{,}1}{3}, y_0 + \tfrac{0{,}1}{3}K_1) \cong 0{,}033333 \\ K_3 &= f(t_0 + \tfrac{0{,}2}{3}, y_0 + \tfrac{0{,}2}{3}K_2) \cong 0{,}066815 \\ y_1 &= y_0 + \tfrac{0{,}1}{4}(K_1 + 3K_3) \cong 1{,}005011 \cong y(0{,}1) \end{aligned}$$

En tanto que para el segundo, partiendo de $t_1 = 0{,}1$ e $y_1 = 1{,}005011$, se tendrá:

$$\begin{aligned} &K_1 = f(t_1, y_1) \cong 0{,}100501 \\ &K_2 = f(t_1 + \tfrac{0{,}1}{3}, y_1 + \tfrac{0{,}1}{3}K_1) \cong 0{,}134448 \\ &K_3 = f(t_1 + \tfrac{0{,}2}{3}, y_1 + \tfrac{0{,}2}{3}K_2) \cong 0{,}1689957 \\ &y_2 = y_1 + \tfrac{0{,}1}{4}(K_1 + 3K_3) \cong \boxed{1{,}020198 \cong y(0{,}2)} \end{aligned}$$

4. Dar el algoritmo del método Runge-Kutta explícito de tres etapas cuyo tablero de Butcher tiene la forma

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 3/4 & 0 & 3/4 & \\ \hline & 2/9 & 3/9 & 4/9 \end{array}$$

¿cuál es su orden?. Aproximar, mediante dicho método, la solución en el tiempo $t = 0{,}4$ del PVI: $y'(t) = t + e^{y(t)}$, $y(0) = 0$, tomando $h = 0{,}2$.

Solución. Al método Runge-Kutta explícito de tres etapas dado por la tabla del enunciado corresponde el algoritmo

$$\begin{aligned} &K_1 = f(t_n, y_n) \\ &K_2 = f(t_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hK_1) \\ &K_3 = f(t_n + \tfrac{3}{4}h, y_n + \tfrac{3}{4}hK_2) \\ &y_{n+1} = y_n + \tfrac{h}{9}(2K_1 + 3K_2 + 4K_3) \end{aligned}$$

Puesto que se cumplen las condiciones de simplificación, veamos las ecuaciones de orden dadas por las fórmulas:

$$\begin{aligned} &1)\ b_1 + b_2 + b_3 = \tfrac{2}{9} + \tfrac{3}{9} + \tfrac{4}{9} = 1 \\ &2)\ b_2c_2 + b_3c_3 = \tfrac{3}{9}\tfrac{1}{2} + \tfrac{4}{9}\tfrac{3}{4} = \tfrac{1}{2} \\ &3)\ b_2c_2^2 + b_3c_3^2 = \tfrac{3}{9}(\tfrac{1}{2})^2 + \tfrac{4}{9}(\tfrac{3}{4})^2 = 1/3 \\ &4)\ b_3a_{32}c_2 = \tfrac{4}{9}\tfrac{3}{4}\tfrac{1}{2} = \tfrac{1}{6} \end{aligned}$$

como se verifican todas el método es de orden tres, el máximo alcanzable por un Runge-Kutta explícito de tres etapas.

Para aproximar la solución del PVI del enunciado en $t = 0{,}4$, partiendo de $t_0 = 0$ e $y_0 = 0$ con paso $h = 0{,}2$, siendo ahora $f(t, y) = t + e^y$, hemos de realizar dos pasos, de manera que reteniendo tan sólo seis cifras decimales significativas, para el primer paso se obtiene:

$$\begin{aligned} &K_1 = f(t_0, y_0) = 1 \\ &K_2 = f(t_0 + \tfrac{1}{2} \cdot 0{,}2, y_0 + \tfrac{1}{2} \cdot 0{,}2K_1) \cong 1{,}205171 \\ &K_3 = f(t_0 + \tfrac{3}{4} \cdot 0{,}2, y_0 + \tfrac{3}{4} \cdot 0{,}2K_2) \cong 1{,}348146 \\ &y_1 = y_0 + \tfrac{0{,}2}{9}(2K_1 + 3K_2 + 4K_3) \cong 0{,}244624 \cong y(0{,}2) \end{aligned}$$

En tanto que para el segundo, partiendo de $t_1 = 0{,}2$ e $y_1 = 0{,}244624$, se tendrá:

$$K_1 = f(t_1, y_1) \cong 1{,}477141$$
$$K_2 = f(t_1 + \tfrac{1}{2} \cdot 0{,}2, y_1 + \tfrac{1}{2} \cdot 0{,}2K_1) \cong 1{,}780438$$
$$K_3 = f(t_1 + \tfrac{3}{4} \cdot 0{,}2, y_1 + \tfrac{3}{4} \cdot 0{,}2K_2) \cong 2{,}018107$$
$$y_2 = y_1 + \tfrac{0{,}2}{9}(2K_1 + 3K_2 + 4K_3) \cong \boxed{0{,}608358 \cong y(0{,}4)}$$

5. Obtener el método Runge-Kutta explícito de tres etapas, cuyo tablero de Butcher tiene la forma

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2 & & \\ 2/3 & a_{31} & a_{32} & \\ \hline & 1/4 & 3/8 & b_3 \end{array}$$

y cuyo orden es máximo. Aproximar, mediante dicho método, la solución en el tiempo $t = 2{,}2$ del PVI: $y' = 1 - \frac{y}{t}$, $y(2) = 2$, tomando $h = 0{,}1$; hacerlo también por el método de Taylor de orden tres.

Solución. En primer lugar, imponemos a los coeficientes verificar la condición de simplificación y las de orden

$$\begin{array}{l} 0)\ \frac{2}{3} = a_{31} + a_{32} \\ 1)\ b_1 + b_2 + b_3 = \frac{1}{4} + \frac{3}{8} + b_3 = 1 \\ 2)\ b_2c_2 + b_3c_3 = \frac{3}{8}c_2 + b_3\frac{2}{3} = \frac{1}{2} \\ 3)\ b_2c_2^2 + b_3c_3^2 = \frac{3}{8}c_2^2 + b_3(\frac{2}{3})^2 = 1/3 \\ 4)\ b_3a_{32}c_2 = \frac{1}{6} \end{array}$$

De la ecuación 1) resulta que $b_3 = \frac{3}{8}$ y llevándolo a la segunda, deducimos que $\frac{3}{8}c_2 = \frac{1}{2} - \frac{3}{8}\frac{2}{3} = \frac{1}{4}$ de donde resulta $c_2 = \frac{8}{3}\frac{1}{4} = \frac{2}{3}$, con estas elecciones de b_3 y c_2, el método tiene orden, al menos dos, por verificar 1) y 2) cualesquiera que sean el resto de coeficientes siempre que se verifique también la 0), veamos si se cumplen las restantes, es inmediato comprobar que con esta elección también se cumple la 3), pues $b_2c_2^2 + b_3c_3^2 = \frac{3}{8}(\frac{2}{3})^2 + \frac{3}{8}(\frac{2}{3})^2 = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$, veamos ahora como elegir a_{32} desde 4) para que el método tenga orden máximo, debe verificarse $b_3a_{32}c_2 = \frac{3}{8}a_{32}\frac{2}{3} = \frac{1}{6}$ de donde se deduce que $a_{32} = \frac{2}{3}$ y de la condición de simplificación 0), resulta $a_{31} = 0$. Así pues, el método Runge-Kutta explícito de orden máximo y de la forma indicada tiene por tablero de Butcher el siguiente:

$$\begin{array}{c|ccc} 0 & & & \\ 2/3 & 2/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 3/8 & 3/8 \end{array}$$

Ahora con el algoritmo correspondiente a este tablero, vamos a aproximar la solución en el tiempo $t = 2{,}2$ del PVI dado, tomando $h = 0{,}1$, partiendo del punto $t_0 = 2$, $y_0 = 2$ y siendo ahora $f(t, y) = 1 - \frac{y}{t}$. Tenemos que hacer dos pasos, y reteniendo tan sólo seis cifras decimales significativas, para el primer paso se obtiene:

$$\begin{aligned}
&K_1 = f(t_0, y_0) = 0\\
&K_2 = f(t_0 + \tfrac{2}{3}\cdot 0{,}1, y_0 + \tfrac{2}{3}\cdot 0{,}1K_1) \cong 0{,}032258\\
&K_3 = f(t_0 + \tfrac{2}{3}\cdot 0{,}1, y_0 + \tfrac{2}{3}\cdot 0{,}1K_2) \cong 0{,}031217\\
&y_1 = y_0 + \tfrac{0{,}1}{8}(2K_1 + 3K_2 + 3K_3) \cong 2{,}002380 \cong y(0{,}1)
\end{aligned}$$

En tanto que para el segundo, partiendo de $t_1 = 2{,}1$ e $y_1 = 2{,}002380$, se tendrá:

$$\begin{aligned}
&K_1 = f(t_1, y_1) \cong 0{,}046486\\
&K_2 = f(t_1 + \tfrac{2}{3}\cdot 0{,}1, y_1 + \tfrac{2}{3}\cdot 0{,}1K_1) \cong 0{,}074394\\
&K_3 = f(t_1 + \tfrac{2}{3}\cdot 0{,}1, y_1 + \tfrac{2}{3}\cdot 0{,}1K_2) \cong 0{,}073536\\
&y_2 = y_1 + \tfrac{0{,}1}{8}(2K_1 + 3K_2 + 3K_3) \cong \boxed{2{,}009090 \cong y(0{,}2)}
\end{aligned}$$

Nos piden hacerlo también por el método de Taylor de orden tres, cuyo algoritmo viene dado por la fórmula:

$$y_{n+1} = y_n + y'_n h + y''_n \frac{h^2}{2} + y'''_n \frac{h^3}{6}$$

siendo

$$\begin{aligned}
&y'_n = f(t_n, y_n) = 1 - \tfrac{y_n}{t_n}\\
&y''_n = f^{(1)}(t_n, y_n) = f_t(t_n, y_n) + f_y(t_n, y_n)\cdot f(t_n, y_n)\\
&y'''_n = f^{(2)}(t_n, y_n) = f_t^{(1)}(t_n, y_n) + f_y^{(1)}(t_n, y_n)\cdot f(t_n, y_n)
\end{aligned}$$

luego se calcularían estas derivadas y se sustituiría en las expresiones anteriores, pero en la práctica resulta más ventajoso hacerlo como sigue, dado que:

$$\begin{aligned}
&y' = f(t, y) = 1 - \tfrac{y}{t}\\
&y'' = -\tfrac{y't - y}{t^2} = -\tfrac{y'}{t} + \tfrac{y}{t^2}\\
&y''' = -\tfrac{y''t - y'}{t^2} + \tfrac{y't^2 - 2ty}{t^4} = -\tfrac{y''}{t} + 2\tfrac{y'}{t^2} - 2\tfrac{y}{t^3}
\end{aligned}$$

entonces particularizando en t_n e indicando $y(t_n)$ por y_n, $y'(t_n)$ por y'_n, etc., resultan

$$\begin{aligned}
&y'_n = 1 - \tfrac{y_n}{t_n}\\
&y''_n = -\tfrac{y'_n}{t_n} + \tfrac{y_n}{t_n^2}\\
&y'''_n = -\tfrac{y''_n}{t_n} + 2\tfrac{y'_n}{t_n^2} - 2\tfrac{y_n}{t_n^3}
\end{aligned}$$

así al calcular y_n'' se pone el valor obtenido de y_n', al calcular y_n''' se ponen los valores de y_n' e y_n'' ya obtenidos. En nuestro caso, partiendo del punto $t_0 = 2$, $y_0 = 2$ y con $h = 0{,}1$, tenemos que hacer dos pasos, y reteniendo tan sólo seis cifras decimales significativas, para el primer paso se tiene:

$$\begin{aligned}
y_0' &= 1 - \tfrac{y_0}{t_0} = 1 - \tfrac{2}{2} = 0\\
y_0'' &= \tfrac{y_0'}{t_0} + \tfrac{y_0}{t_0^2} = 0 + \tfrac{2}{2^2} = \tfrac{1}{2}\\
y_0''' &= -\tfrac{y_0''}{t_0} + 2\tfrac{y_0'}{t_0^2} - 2\tfrac{y_0}{t_0^3} = -\tfrac{1}{4} + 0 - \tfrac{1}{2} = -\tfrac{3}{4}\\
y_1 &= y_0 + y_0'h + y_0''\tfrac{h^2}{2} + y_0'''\tfrac{h^3}{6} = 2{,}002375
\end{aligned}$$

y para el segundo, partiendo de $t_1 = 2{,}1$ e $y_1 = 2{,}002375$ con $h = 0{,}1$, resulta:

$$\begin{aligned}
y_1' &= 1 - \tfrac{y_1}{t_1} \cong 0{,}046488\\
y_1'' &= \tfrac{y_1'}{t_1} + \tfrac{y_1}{t_1^2} \cong 0{,}431916\\
y_1''' &= -\tfrac{y_1''}{t_1} + 2\tfrac{y_1'}{t_1^2} - 2\tfrac{y_1}{t_1^3} \cong -0{,}617023\\
y_2 &= y_1 + y_1'h + y_1''\tfrac{h^2}{2} + y_1'''\tfrac{h^3}{6} \cong \boxed{2{,}009081 \cong y(2{,}2)}
\end{aligned}$$

6. Hallar el método Runge-Kutta explícito de tres etapas y orden máximo, cuyo tablero de Butcher tiene la forma

$$\begin{array}{c|ccc}
0 & & & \\
1/2 & 1/2 & & \\
1 & a_{31} & a_{32} & \\
\hline
 & b_1 & 2/3 & b_3
\end{array}$$

Aproximar la solución en el tiempo $t = 0{,}2$ del PVI $y'(t) = t^2 + 2y$, $y(0) = 1$; tomando $h = 0{,}1$ e integrándolo mediante el método hallado en el apartado anterior. Asimismo, tras integrar por dicho método en un sólo paso de amplitud $h = 0{,}2$, estimar, por el método de extrapolación de Richardson, el error cometido en la primera aproximación de $y(0{,}2)$ hallada y mejorar dicha aproximación.

Solución. Para comenzar, imponemos a los coeficientes verificar la condición de simplificación que falta y las condiciones de orden

$$\begin{aligned}
&0)\ 1 = a_{31} + a_{32}\\
&1)\ b_1 + b_2 + b_3 = b_1 + \tfrac{2}{3} + b_3 = 1\\
&2)\ b_2c_2 + b_3c_3 = \tfrac{2}{3}\tfrac{1}{2} + b_3 \cdot 1 = \tfrac{1}{2}\\
&3)\ b_2c_2^2 + b_3c_3^2 = \tfrac{2}{3}(\tfrac{1}{2})^2 + b_3(1)^2 = 1/3\\
&4)\ b_3a_{32}c_2 = b_3a_{32}\tfrac{1}{2} = \tfrac{1}{6}
\end{aligned}$$

De la ecuación 2) se deduce que $b_3 = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$ y de la primera que $b_1 = 1 - \frac{1}{6} - \frac{2}{3} = \frac{1}{6}$, verificándose 0), 1) y 2) el método es de orden, al menos, dos, veamos si se cumplen las restantes 3) $b_2c_2^2 + b_3c_3^2 = \frac{2}{3}(\frac{1}{2})^2 + \frac{1}{6}(1)^2 = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$, que también se cumple y la 4) se cumple si se toma $a_{32} = 2$, con lo cual habrá de tomarse $a_{31} = -1$ para que también se verifique la 0); por tanto el método de orden 3 (máximo alcanzable) de esta forma será el dado por el tablero siguiente:

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 4/6 & 1/6 \end{array}$$

Hemos de realizar en primer lugar dos pasos, con $h = 0{,}1$, partiendo de $t_0 = 0$, $y_0 = 1$ y siendo $f(t, y) = t^2 + 2y$, reteniendo como en los anteriores tan sólo seis cifras decimales significativas, para el primer paso se obtiene:

$$\begin{array}{l} K_1 = f(t_0, y_0) = 2 \\ K_2 = f(t_0 + \frac{1}{2} \cdot 0{,}1, y_0 + \frac{1}{2} \cdot 0{,}1K_1) = 2{,}2025 \\ K_3 = f(t_0 + 0{,}1, y_0 + 0{,}1(-K_1 + 2K_2)) = 2{,}491 \\ y_1 = y_0 + \frac{0{,}1}{6}(K_1 + 4K_2 + K_3) \cong 1{,}221683 \cong y(0{,}1) \end{array}$$

En tanto que para el segundo, partiendo de $t_1 = 0{,}1$ e $y_1 = 1{,}221683$, se tendrá:

$$\begin{array}{l} K_1 = f(t_1, y_1) = 2{,}453366 \\ K_2 = f(t_1 + \frac{1}{2} \cdot 0{,}1, y_1 + \frac{1}{2} \cdot 0{,}1K_1) \cong 2{,}711203 \\ K_3 = f(t_1 + 0{,}1, y_1 + 0{,}1(-K_1 + 2K_2)) \cong 3{,}077174 \\ y_2 = y_1 + \frac{0{,}1}{6}(K_1 + 4K_2 + K_3) \boxed{\cong 1{,}494606 \cong y(0{,}2)} \end{array}$$

Ahora aproximaremos $y(0{,}2)$ con un único paso de amplitud $h = 0{,}2$, partiendo de $t_0 = 0$, $y_0 = 1$. resultando

$$\begin{array}{l} K_1 = f(t_0, y_0) = 2 \\ K_2 = f(t_0 + \frac{1}{2} \cdot 0{,}1, y_0 + \frac{1}{2} \cdot 0{,}1K_1) = 2{,}41 \\ K_3 = f(t_0 + 0{,}1, y_0 + 0{,}1(-K_1 + 2K_2)) = 3{,}168 \\ y_1 = y_0 + \frac{0{,}1}{6}(K_1 + 4K_2 + K_3) = \boxed{1{,}4936 \cong y(0{,}2)} \end{array}$$

Ahora se puede estimar el error global de la primera aproximación hallada para $y(0{,}2)$ por el método de extrapolación al límite de Richardson, mediante la fórmula

$$y(0{,}2) - y_{0,1}(0{,}2) \cong \frac{y_{0,1}(0{,}2) - y_{0,2}(0{,}2)}{2^p - 1} = E$$

donde $y(0{,}2)$ es el valor de la solución exacta en el punto $t = 0{,}2$, $y_{0,1}(0{,}2)$ es la aproximación numérica de la solución en dicho punto, obtenida con el método numérico dado, de orden $p = 3$, con el paso 0,1, e $y_{0,2}(0{,}2)$ la obtenida con paso $h = 0{,}2$, en este caso el orden del método es $p = 3$, resultando la estimación

$$E = \frac{1{,}494606 - 1{,}4936}{2^3 - 1} \cong \boxed{1{,}437143 \cdot 10^{-4}}$$

La aproximación mejorada de $y(0{,}2)$, por el método de extrapolación al límite de Richardson se obtiene sumando a $y_{0,1}(0{,}2)$ la estimación del error E, en este caso resulta ser

$$AM = y_{0,1}(0{,}2) + E \cong \boxed{1{,}4947497}$$

(Con objeto de que podamos comparar las aproximaciones y estimaciones obtenidas, consignemos la solución exacta que viene dada por $y(0{,}2) = 1{,}494780872...$).

7. Estudiar en función de c_3 y b_2 el orden del método Runge-Kutta explícito de tres etapas cuyo tablero de Butcher tenga la forma

$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ c_3 & 0 & c_3 & \\ \hline & 1/4 & b_2 & 3/4 \end{array}$$

y obtener uno de orden tres. Aproximar la solución en el tiempo $t = 0{,}2$ del PVI $y'(t) = t^2 - y$ con la condición inicial $y(0) = 1$, por dicho método mediante un sólo paso de amplitud $h = 0{,}2$, hacerlo también mediante dos pasos de amplitud $h = 0{,}1$. Asimismo, mejorar la última aproximación obtenida por el método de Richardson. Ahora, teniendo en cuenta que la solución exacta está dada por $y(t) = -e^{-t} + t^2 - 2t + 2$, decir si los errores globales en $t = 0{,}2$ se comportan como era de esperar cuando h se divide por dos. Finalmente, utilizar el método Runge-Kutta clásico de cuarto orden para aproximar $y(0{,}2)$, realizando dos pasos de amplitud $h = 0{,}1$.

Solución. Puesto que se cumplen las condiciones de simplificación, las condiciones de orden para los RK(3) que tienen este tablero de Butcher son:

1) $b_1 + b_2 + b_3 = \frac{1}{4} + b_2 + \frac{3}{4} = 1$
2) $b_2 c_2 + b_3 c_3 = b_2 \frac{1}{3} + \frac{3}{4} c_3 = \frac{1}{2}$
3) $b_2 c_2^2 + b_3 c_3^2 = b_2 (\frac{1}{3})^2 + \frac{3}{4} c_3^2 = 1/3$
4) $b_3 a_{32} c_2 = \frac{3}{4} c_3 \frac{1}{3} = \frac{1}{6}$

Para orden, al menos, uno se deduce de 1) que debe ser $b_2 = 0$, en tanto que de la 2) deducimos que $c_3 = \frac{2}{3}$, con estos dos valores se puede comprobar que se cumplen la 3) y 4); por tanto podemos concluir que si $b_2 = 0$ el método es convergente de orden, al menos, uno y que si además es $c_3 = \frac{2}{3}$ el orden es tres, pero si $c_3 \neq \frac{2}{3}$ sería de primer orden. Con el método de orden tres obtenido demos un paso de amplitud $h = 0{,}2$, partiendo de $t_0 = 0$, $y_0 = 1$ y siendo $f(t, y) = t^2 - y$, entonces resulta

$$\begin{array}{l} K_1 = f(t_0, y_0) = -1 \\ K_2 = f(t_0 + \frac{1}{3} \cdot 0{,}2, y_0 + \frac{1}{3} \cdot 0{,}2K_1) = -0{,}92888888888889 \\ K_3 = f(t_0 + \frac{2}{3} \cdot 0{,}2, y_0 + \frac{2}{3} \cdot 0{,}2K_2) = -0{,}85837037037037 \\ y_1 = y_0 + \frac{0{,}2}{4}(K_1 + 3K_3) = \boxed{0{,}821244 \cong y(0{,}2)} \end{array}$$

Ahora haremos dos pasos de amplitud $h = 0{,}1$ partiendo del mismo punto, para el primero, obtendremos:

$$\begin{array}{l} K_1 = f(t_0, y_0) = -1 \\ K_2 = f(t_0 + \frac{1}{3} \cdot 0{,}1, y_0 + \frac{1}{3} \cdot 0{,}1K_1) = -0{,}96555555555556 \\ K_3 = f(t_0 + \frac{2}{3} \cdot 0{,}1, y_0 + \frac{2}{3} \cdot 0{,}1K_2) = -0{,}93118518518519 \\ y_1 = y_0 + \frac{0{,}2}{4}(K_1 + 3K_3) = 0{,}905161 \cong y(0{,}1) \end{array}$$

y para el segundo, partiendo de $t_1 = 0{,}1$ e $y_1 = 0{,}905161$, resulta:

$$\begin{array}{l} K_1 = f(t_1, y_1) = -0{,}895161 \\ K_2 = f(t_1 + \frac{1}{3} \cdot 0{,}1, y_1 + \frac{1}{3} \cdot 0{,}1K_1) = -0{,}85754452222222 \\ K_3 = f(t_1 + \frac{2}{3} \cdot 0{,}1, y_1 + \frac{2}{3} \cdot 0{,}1K_2) = -0{,}82021358740741 \\ y_2 = y_1 + \frac{0{,}1}{4}(K_1 + 3K_3) \cong \boxed{0{,}821266 \cong y(0{,}2)} \end{array}$$

Estimemos el error de esta segunda aproximación de $y(0{,}2)$ mediante el método de Richardson, se tendrá, de acuerdo con lo expuesto en el ejercicio anterior

$$E = \frac{0{,}821266 - 0{,}821244}{2^3 - 1} \cong \boxed{3{,}14 \cdot 10^{-6}}$$

La aproximación mejorada de $y(0{,}2)$, por el método de extrapolación al límite de Richardson se obtiene sumando a $y_{0,1}(0{,}2)$ la estimación del error E, en este caso resulta ser

$$AM = y_{0,1}(0{,}2) + E = 0{,}821266 + 3{,}14 \cdot 10^{-6} \cong \boxed{0{,}821269}$$

El valor exacto de la solución en $t = 0{,}2$, calculado con la solución dada, es $y(0{,}2) = 0{,}821269246922...$. Ahora los errores globales de las

aproximaciones obtenidas con pasos $h = 0{,}2$ y $h = 0{,}1$ serán

$$EG1 = y(0{,}2) - y_{0,2}(0{,}2) \cong 0{,}821269 - 0{,}821244 = 0{,}000025$$

$$EG2 = y(0{,}2) - y_{0,1}(0{,}2) \cong 0{,}821269 - 0{,}821266 = 0{,}000003$$

como se observa al dividir el paso por 2 el error se divide aproximadamente por 2^3, como era de esperar en un método de orden 3; también se observa la coincidencia de las seis primeras cifras de la aproximación mejorada con las de la solución exacta.

Recordemos que el algoritmo del método Runge-Kutta clásico de cuatro etapas y orden cuatro tiene el siguiente tablero de Butcher:

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array}$$

Utilizando este método realizaremos dos iteraciones, para aproximar $y(0{,}2)$, con paso $h = 0{,}1$, partiendo de $t_0 = 0$ e $y_0 = 1$, entonces obtendremos tras el primer paso:

$$\begin{aligned}
&K_1 = f(t_0, y_0) = -1 \\
&K_2 = f(t_0 + \tfrac{1}{2} \cdot 0{,}1, y_0 + \tfrac{1}{2} \cdot 0{,}1K_1) = -0{,}9475 \\
&K_3 = f(t_0 + \tfrac{1}{2} \cdot 0{,}1, y_0 + \tfrac{1}{2} \cdot 0{,}1K_2) = -0{,}950125 \\
&K_4 = f(t_0 + 0{,}1, y_0 + 0{,}1K_3) = -0{,}8949875 \\
&y_1 = y_0 + \tfrac{0{,}1}{6}(K_1 + 2(K_2 + K_3) + K_4) \cong 0{,}9051627 \cong y(0{,}1)
\end{aligned}$$

donde hemos retenido en y_1 sólo siete cifras decimales, en tanto que para el segundo partiendo de $t_1 = 0{,}1$ e $y_1 = 0{,}9051627$, resulta

$$\begin{aligned}
&K_1 = f(t_1, y_1) \cong -0{,}8951627 \\
&K_2 = f(t_1 + \tfrac{1}{2} \cdot 0{,}1, y_1 + \tfrac{1}{3} \cdot 0{,}1K_1) \cong -0{,}837904565 \\
&K_3 = f(t_1 + \tfrac{2}{3} \cdot 0{,}1, y_1 + \tfrac{2}{3} \cdot 0{,}1K_2) \cong -0{,}84076747175 \\
&K_4 = f(t_0 + 0{,}1, y_0 + 0{,}1K_3) \cong -0{,}781085952825 \\
&y_2 = y_1 + \tfrac{0{,}1}{6}(K_1 + 2(K_2 + K_3) + K_4) = \boxed{0{,}8212695 \cong y(0{,}2)}
\end{aligned}$$

aproximación cuyas seis primeras cifras decimales coinciden con la solución exacta.

8. Se quiere aproximar la solución en $t = 2{,}04$ del problema

$$y' = t + \sin y, \quad y(2) = 0{,}3$$

a) Obtén el método de Adams-Bashforth explícito de $k = 3$ pasos.

b) Aproxima $y(2{,}01)$ e $y(2{,}02)$ utilizando el método de Taylor de orden adecuado para a continuación realizar el apartado c) que sigue.

c) A partir de los datos anteriores aproxima $y(2{,}03)$ e $y(2{,}04)$ con el método de Adams-Bashforth obtenido en el apartado a) anterior.

Solución. a) La fórmula de Adams-Bashforth de $k = 3$ pasos, abreviadamente AB(3), puede escribirse en la forma

$$\begin{cases} y_n = y_{n-1} + h(\gamma_0 y'_{n-1} + \gamma_1 \nabla y'_{n-1} + \gamma_2 \nabla^2 y'_{n-1}) \\ y'_n = f(t_n, y_n) \end{cases}$$

y siendo

$$\gamma_j = (-1)^j \int_0^1 \binom{-\tau}{j} d\tau$$

resultan

$$\gamma_0 = \int_0^1 d\tau = 1 \quad \gamma_1 = \int_0^1 \tau d\tau = \frac{1}{2} \;\text{ y }\; \gamma_2 = \int_0^1 \frac{\tau(\tau+1)}{2} d\tau = \frac{5}{12}$$

ahora, de las fórmulas

$$\nabla y'_{n-1} = y'_{n-1} - y'_{n-2}$$

$$\nabla^2 y'_{n-1} = \nabla y'_{n-1} - \nabla y'_{n-2} = y'_{n-1} - y'_{n-2} - (y'_{n-2} - y'_{n-3}) = y'_{n-1} - 2y'_{n-2} + y'_{n-3}$$

sustituyendo en la anterior nos queda el método AB(3) en la forma:

$$\begin{cases} y_n = y_{n-1} + \frac{h}{12}(23y'_{n-1} - 16y'_{n-2} + 5y'_{n-3}) \\ y'_n = f(t_n, y_n) \end{cases}$$

como sabemos dicho método es de tercer orden.

b) Para inicializar el método de Adams-Bashforth de 3 pasos y orden 3, necesitamos tres valores $y(2{,}0), y(2{,}01)$ e $y(2{,}02)$, luego hemos de utilizar un método de un paso de orden, al menos, 3, para generar aproximaciones a $y(2{,}01)$ e $y(2{,}02)$ con las que poder iniciar el AB(3). Nos piden que utilicemos el método de Taylor de orden adecuado, utilizaremos por tanto el de orden tres, cuyo algoritmo puede escribirse en la forma:

$$y_{n+1} = y_n + y'_n h + y''_n \frac{h^2}{2} + y'''_n \frac{h^3}{6}$$

siendo

$$\begin{aligned} y'_n &= t_n + \sin y_n \\ y''_n &= 1 + y'_n \cos y_n \\ y'''_n &= y''_n \cos y_n - (y'_n)^2 \sin y_n \end{aligned}$$

Ahora, partiendo del punto $t_0 = 2$, $y_0 = 0{,}3$ con $h = 0{,}01$, tenemos que hacer dos pasos, y reteniendo en los cálculos ocho cifras decimales significativas, para el primer paso se tiene:

$$\begin{aligned} y'_0 &= 2{,}295520206661339 \\ y''_0 &= 3{,}19299421494873 \\ y'''_0 &= 1{,}493165858692355 \\ y_1 = y_0 + y'_0 h + y''_0 \tfrac{h^2}{2} + y'''_0 \tfrac{h^3}{6} &\cong \boxed{0{,}32311510 \cong y(2{,}01)} \end{aligned}$$

y para el segundo, partiendo de $t_1 = 2{,}01$ e $y_1 = 0{,}32311510$ con el mismo paso, resulta:

$$\begin{aligned} y'_1 &\cong 2{,}327521992832856 \\ y''_1 &\cong 3{,}207074842534859 \\ y'''_1 &\cong 1{,}320981123513974 \\ y_2 = y_1 + y'_1 h + y''_1 \tfrac{h^2}{2} + y'''_1 \tfrac{h^3}{6} &\cong \boxed{0{,}34655089 \cong y(2{,}02)} \end{aligned}$$

c) Utilizando los datos anteriores vamos a obtener, mediante el AB(3) aproximaciones de la solución en los puntos 2,03 y 2,04, resultando para la primera

$$\begin{cases} y_3 = y_2 + \frac{h}{12}(23y'_2 - 16y'_1 + 5y'_0) \\ y'_3 = f(t_3, y_3) \end{cases}$$

puesto que ya hemos calculado y_2, y'_0 e y'_1, nos falta y'_2 para aplicar la AB(3) e y'_3 para el paso siguiente, que resultan ser

$$\begin{aligned} & y'_2 = f(t_2, y_2) = t_2 + \sin(y_2) = 2{,}02 + \sin(0{,}34655089) = 2{,}35965577444605 \\ & y_3 = 0{,}34655089 + \tfrac{0{,}01}{12}(23 \cdot 2{,}35965577444605 - 16 \cdot 2{,}327521992832856 + \\ & +5 \cdot 2{,}295520206661339) = \boxed{0{,}37030867 \cong y(2{,}03)} \\ & y'_3 = f(t_3, y_3) = t_3 + \sin(y_3) = 2{,}391903193077037 \end{aligned}$$

y para la segunda

$$y_4 = y_3 + \tfrac{h}{12}(23y'_3 - 16y'_2 + 5y'_1) \cong \boxed{0{,}39438941 \cong y(2{,}04)}$$

9. Se quiere aproximar la solución del PVI $y'(t) = t^2 + 2y$, $y(0) = 1$; en el punto $t = 0{,}3$. Para ello, se pide:

a) Obtener el método convergente de mayor orden de la forma

$$y_n - y_{n-2} = h[\beta_2 f_n + \beta_1 f_{n-1} + \beta_0 f_{n-2}]$$

b) Aproximar con el método de Taylor del mismo orden que el MLM obtenido en el apartado anterior $y(0{,}1)$.

c) A partir de los datos $y(0)$ e $y(0{,}1)$ aproximar $y(0{,}2)$ e $y(0{,}3)$ con el método obtenido en el apartado a) anterior.

Solución. a) Se trata de un método lineal multipaso implícito de dos pasos, puesto que su primer polinomio característico es λ^2-1, sus raíces son $\lambda_1 = 1$ y $\lambda_2 = -1$, que aunque tienen módulo 1 son simples, por tanto el método es estable, luego será convergente si su orden es mayor o igual que uno; dado que los coeficientes del primer polinomio son $\alpha_2 = 1$, $\alpha_1 = 0$, $alpha_0 = -1$ y los del segundo polinomio característico son β_2, β_1 y β_0, las ecuaciones de orden (teniendo en cuenta que el orden máximo alcanzable es 4 por ser un método implícito de dos pasos) son:

$$\begin{array}{l}
c_0 = \sum_{j=0}^{2} \alpha_j = 1 + 0 - 1 = 0 \\
c_1 = \sum_{j=0}^{2} (j\alpha_j - \beta_j) = 0 \Leftrightarrow \beta_2 + \beta_1 + \beta_0 = 2 \\
c_2 = \frac{1}{2!} \sum_{j=0}^{2} (j^2\alpha_j - 2j\beta_j) = 0 \Leftrightarrow 2\beta_2 + \beta_1 = 2 \\
c_3 = \frac{1}{3!} \sum_{j=0}^{2} (j^3\alpha_j - 3j^2\beta_j) = 0 \Leftrightarrow 2^2\beta_2 + \beta_1 = \frac{8}{3} \\
c_4 = \frac{1}{4!} \sum_{j=0}^{2} (j^4\alpha_j - 4j^3\beta_j) = 0 \Leftrightarrow 2^3\beta_2 + \beta_1 = 4
\end{array}$$

Para que sea de orden, al menos, 1 deben verificarse las dos primeras ecuaciones, de orden al menos, dos las tres primeras, para orden, al menos, tres las cuatro primeras, de las que se deduce que $\beta_0 = \frac{1}{3}$, $\beta_1 = \frac{4}{3}$ y $\beta_2 = \frac{1}{3}$, con estos también se verifica que $c_4 = 0$, luego para esos valores el método es convergente de orden cuatro (el máximo alcanzable).

b) Utilizaremos como nos piden el método de Taylor de orden cuatro para aproximar $y(0{,}1)$, el algoritmo del mismo viene dado por

$$y_{n+1} = y_n + y'_n h + y''_n \frac{h^2}{2} + y'''_n \frac{h^3}{6} + y_n^{(iv)} \frac{h^4}{24}$$

siendo

$$\begin{array}{l}
y'_n = t_n^2 + 2y_n \\
y''_n = 2t_n + 2y'_n \\
y'''_n = 2 + 2y''_n \\
y_n^{(iv)} = 2y'''_n
\end{array}$$

Ahora, partiendo del punto $t_0 = 0$, $y_0 = 1$ daremos un único paso de amplitud $h = 0{,}1$, obteniéndose:

$$\begin{array}{l} y_0' = t_0^2 + 2y_0 = 2 \\ y_0'' = 2t_0 + 2y_0' = 4 \\ y_0''' = 2 + 2y_0'' = 10 \\ y_0^{(iv)} = 2y_0''' = 20 \\ y_1 = y_0 + y_0'h + y_0''\frac{h^2}{2} + y_0'''\frac{h^3}{6} + y_0^{(iv)}\frac{h^4}{24} = \boxed{1{,}22175 \cong y(0{,}1)} \end{array}$$

c) En este apartado, hemos de obtener aproximaciones de $y(0{,}2)$ e $y(0{,}3)$ partiendo del $y(0)$ dado y del $y(0{,}1)$ obtenido en el apartado anterior, mediante el MLM obtenido en a)con paso $h = 0{,}1$, dicho método está dado por el algoritmo

$$y_n - y_{n-2} = \frac{h}{3}[f_n + 4f_{n-1} + f_{n-2}]$$

resultando para $n = 2$

$$y_2 - y_0 = \frac{0{,}1}{3}[f_2 + 4f_1 + f_0]$$

ahora puesto que $f_2 = f(t_2, y_2) = t_2^2 + 2y_2$, $f_1 = f(t_1, y_1) = t_1^2 + 2y_1 = 2{,}4535$ y $f_0 = f(t_0, y_0) = 2$, sustituyendo en la anterior se tiene

$$y_2 - 1 = \frac{0{,}1}{3}[0{,}2^2 + 2y_2 + 4 \cdot 2{,}4535 + 2]$$

pero como f es lineal en y, se puede despejar y_2 y se obtiene

$$y_2 = [1 + \frac{0{,}1}{3}(0{,}2^2 + 4 \cdot 2{,}4535 + 2)]/(1 - 0{,}2/3) = \boxed{1{,}494785714285714 \cong y(0{,}2)}$$

para dar otro paso, calculemos primero $f_2 = f(t_2, y_2) = 0{,}2^2 + 2y_2 = 3{,}029571428571428$, entonces para $n = 3$, se tendrá

$$y_3 - y_1 = \frac{0{,}1}{3}[f_3 + 4f_2 + f_1]$$

y sustituyendo $f_3 = t_3^2 + 2y_3 = 0{,}3^2 + 2y_3$, así como el resto de términos por sus valores ya obtenidos, podemos nuevamente despejar y_3 en la forma

$$y_3 = [1{,}22175 + \frac{0{,}1}{3}(0{,}3^2 + 4 \cdot 3{,}029571428571428 + 2{,}4535)]/(1 - 0{,}2/3) =$$

$$= \boxed{1{,}83265306122449 \cong y(0{,}3)}$$

10. Se quiere aproximar la solución del PVI: $y'(t) = t^2 + ty$, $y(0) = 1$, en el punto $t = 0{,}4$. Para ello, se pide:

    a) Obtener el método de Adams-Moulton de tres pasos, ¿es estable?, ¿cuál es su orden?.

    b) Aproximar con el método de Taylor del mismo orden que el MLM obtenido en el apartado anterior $y(0{,}1)$ e $y(0{,}2)$.

    c) A partir de $y(0)$ y las aproximaciones de $y(0{,}1)$ e $y(0{,}2)$ obtenidas, aproximar $y(0{,}3)$ e $y(0{,}4)$ con el método del apartado a) anterior.

Solución. a) La fórmula de Adams-Moulton de $k = 3$ pasos, abreviadamente AM(3), puede escribirse en la forma

$$\left\{ \begin{array}{l} y_n = y_{n-1} + h \sum_{j=0}^{3} \gamma_j^* \nabla^j y'_n \\ y'_n = f(t_n, y_n) \end{array} \right.$$

y siendo

$$\gamma_j^* = (-1)^j \int_{-1}^{0} \binom{-\tau}{j} d\tau$$

resultan

$$\gamma_0^* = \int_{-1}^{0} d\tau = 1, \quad \gamma_1^* = \int_{-1}^{0} \tau d\tau = -\frac{1}{2}$$

$$\gamma_2^* = \int_{-1}^{0} \frac{\tau(\tau+1)}{2} d\tau = -\frac{1}{12} \;\text{ y }\; \gamma_3^* = \int_{-1}^{0} \frac{\tau(\tau+1)(\tau+2)}{6} d\tau = -\frac{1}{24}$$

ahora, de las fórmulas

$$\nabla y'_n = y'_n - y'_{n-1}, \quad \nabla^2 y'_n = \nabla y'_n - \nabla y'_{n-1} = y'_n - 2y'_{n-1} + y'_{n-2}$$

$$\text{y} \quad \nabla^3 y'_n = \nabla^2 y'_n - \nabla^2 y'_{n-1} = y'_n - 3y'_{n-1} + 3y'_{n-2} - y'_{n-3}$$

sustituyendo en la anterior nos queda el método AM(3) en la forma:

$$\left\{ \begin{array}{l} y_n = y_{n-1} + \frac{h}{24}(9y'_n + 19y'_{n-1} - 5y'_{n-2} + y'_{n-3}) \\ y'_n = f(t_n, y_n) \end{array} \right.$$

puesto que su primer polinomio característico es $\lambda^3 - \lambda^2$, sus raíces son $\lambda_1 = 0$ doble y $\lambda_2 = 1$ por tanto cumple la condición de las raíces y es estable;como sabemos por teoría dicho método es un MLM implícito de 3 pasos y de cuarto orden, pero puede comprobarse con las ecuaciones de orden.

b) Ahora, debemos aproximar con el método de Taylor de orden cuatro $y(0{,}1)$ e $y(0{,}2)$, el algoritmo correspondiente será

$$y_{n+1} = y_n + y'_n h + y''_n \frac{h^2}{2} + y'''_n \frac{h^3}{6} + y_n^{(iv)} \frac{h^4}{24}$$

siendo

$$\begin{aligned} y'_n &= t_n^2 + t_n y_n \\ y''_n &= 2t_n + y_n + t_n y'_n \\ y'''_n &= 2 + 2y'_n + t_n y''_n \\ y_n^{(iv)} &= 3y''_n + t_n y'''_n \end{aligned}$$

Daremos dos pasos de amplitud $h = 0{,}1$, partiendo del punto $t_0 = 0$, $y_0 = 1$, reteniendo sólo ocho cifras decimales significativas en las aproximaciones pedidas, se obtiene para la primera:

$$\begin{aligned} & y'_0 = t_0^2 + t_0 y_0 = 0 \\ & y''_0 = 2t_0 + y_0 + t_0 y'_0 = 1 \\ & y'''_0 = 2 + 2y'_0 + t_0 y''_0 = 2 \\ & y_0^{(iv)} = 3y''_0 + t_0 y'''_0 = 3 \\ & y_1 = y_0 + y'_0 h + y''_0 \tfrac{h^2}{2} + y'''_0 \tfrac{h^3}{6} + y_0^{(iv)} \tfrac{h^4}{24} \cong \boxed{1{,}00534583 \cong y(0{,}1)} \end{aligned}$$

y para la segunda, partiendo de $t_1 = 0{,}1$, $y_1 = 1{,}00534583$ con paso $h = 0{,}1$, obtendremos:

$$\begin{aligned} & y'_1 = t_1^2 + t_1 y_1 = 0{,}110534583 \\ & y''_1 = 2t_1 + y_1 + t_1 y'_1 = 1{,}2163992883 \\ & y'''_1 = 2 + 2y'_1 + t_1 y''_1 = 2{,}34270909483 \\ & y_1^{(iv)} = 3y''_1 + t_1 y'''_1 = 3{,}883468774383 \\ & y_2 = y_1 + y'_1 h + y''_1 \tfrac{h^2}{2} + y'''_1 \tfrac{h^3}{6} + y_1^{(iv)} \tfrac{h^4}{24} \cong \boxed{1{,}02288792 \cong y(0{,}2)} \end{aligned}$$

c) Ahora, hemos de obtener aproximaciones de $y(0{,}3)$ e $y(0{,}4)$ partiendo del $y(0)$ dado y de las aproximaciones obtenidas en b) de $y(0{,}1)$ e $y(0{,}2)$, realizando dos pasos mediante el AM(3) obtenido en a) con amplitud de paso $h = 0{,}1$; entonces, para el primer paso se tendrá

$$y_3 = y_2 + \tfrac{h}{24}(9y'_3 + 19y'_2 - 5y'_1 + y'_0)$$

y sustituyendo en la anterior $y'_3 = f(t_3, y_3) = t_3^2 + t_3 y_3$, $y'_2 = f(t_2, y_2) = 0{,}2^2 + 0{,}2 \cdot 1{,}02288792 = 0{,}244577584$, $y'_1 = f(t_1, y_1) = 0{,}110534583$ e $y'_0 = f(t_0, y_0) = 0$, nos queda

$$y_3 = 1{,}02288792 + \frac{0{,}1}{24}(9 \cdot (0{,}3^2 + 0{,}3y_3) + 19 \cdot 0{,}244577584 - 5 \cdot 0{,}110534583 + 0)$$

de donde, fácilmente, se despeja y_3 por ser f lineal en y, que resulta ser

$$y_3 = \frac{1{,}02288792 + \frac{0{,}1}{24}(9 \cdot 0{,}3^2 + 19 \cdot 0{,}244577584 - 5 \cdot 0{,}110534583 + 0)}{1 - \frac{0{,}1}{24} \cdot 9 \cdot 0{,}3}$$

$$\cong \boxed{1{,}05519343 \cong y(0{,}3)}$$

análogamente se puede obtener y_4 en la forma

$$y_4 = y_3 + \frac{h}{24}(9y_4' + 19y_3' - 5y_2' + y_1')$$

y sustituyendo como antes se obtiene

$$y_4 = \frac{1{,}05519343 + \frac{0{,}1}{24}(9 \cdot 0{,}4^2 + 19 \cdot 0{,}406558029 - 5 \cdot 0{,}244577584 + 0{,}110534583)}{1 - \frac{0{,}1}{24} \cdot 9 \cdot 0{,}4}$$

$$\cong \boxed{1{,}10532433 \cong y(0{,}4)}$$

6.9. Algunos programas Maxima para métodos de un paso

6.9.1. Métodos de Taylor

A modo de ejemplo, en los diversos métodos que presentaremos, vamos a integrar numéricamente el PVI: $y' = y, y(0) = 1$, en el intervalo $[0, 1]$, para este problema conocemos su solución exacta dada por $y = e^t$.

Comenzaremos con el método de Taylor de orden 3, tomando $h = 0{,}1$. Luego, comparamos la solución obtenida con la exacta dada, mediante las gráficas correspondientes. Para ello, construiremos una función definida por un block, que se puede extender fácilmente para otros PVI y para otros órdenes de convergencia. Con N indicamos el número de subintervalos en el que dividimos el intervalo de integración dado, con $T = b - a$ la amplitud de dicho intervalo (1 en el ejemplo), $h = T/N$ es el paso fijo de integración, $t[0]$ e $y[0]$ son el tiempo inicial y el valor inicial de la solución, $f(t, y)$ es la función que define la ecuación diferencial, en tanto que $f1(t, y)$ y $f2(t, y)$ son respectivamente las derivadas primera y segunda de $f(t, y)$ a lo largo de la solución del problema.

```
(%i1) Taylor3(f,t0,y0,N,T):=block([numer],numer:true,kill(y,t),

      /*Tamaño del paso y valores iniciales*/

      h:T/N, t[0]:t0, y[0]:y0,

      /*Programa del método*/

      define(f1(t,y),diff(f(t,y),t)+diff(f(t,y),y)*f(t,y)),
      define(f2(t,y),(diff(f1(t,y),t)+diff(f1(t,y),y)*f(t,y))),
      define(fi(t,y),f(t,y)+(h/2)*f1(t,y)+(h^2/6)*f2(t,y)),
      for k:1 thru N do (t[k]:t[0]+k*h,y[k]:y[k-1]+h*fi(t[k-1],
      y[k-1])),

      /*Salidas numérica y gráfica.
        Comparación con la solución exacta*/

      print("Solución numérica: ",sol:makelist([t[j],y[j]],j,0,N)),
      wxplot2d([[discrete,sol],%e^t],[t,0,1],[style,[points,2,2],
      [lines,1,1]],[legend,"y(t) aproximada","y(t) exacta"],
      [xlabel,"t"], [ylabel,"y"]),print("El error global en 1.0
      es EG(1.0) = ",%e-y[N]))$

(%i2) f(t,y):=y$ Taylor3(f,0,1,10,1)$
```

Solución numérica: [[0,1],[0.1,1.105166666666667],[0.2,1.221393361111111],
[0.3,1.349843229587963],[0.4,1.491801742566297],[0.5,1.648689559159519],
[0.6,1.822076744464462],[0.7,2.013698482090641],[0.8,2.22547243912384],
[0.9,2.459517957305031],[1.0,2.71817726248161]]

(%t3)

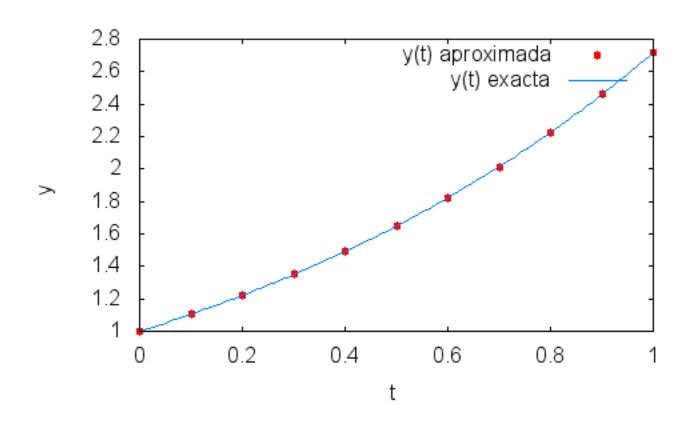


Figura 6.1: Exacta vs aproximada por Taylor de orden 3 y h = 0.1

(%o3)
El error global en 1,0 es $EG(1{,}0) = 1{,}0456597743546681 \cdot 10^{-4}$

Si queremos integrar el mismo PVI con paso $h = 0{,}01$, basta con poner $N = 100$, omitiremos en la salida la solución numérica y sólo pondremos la gráfica, así como el error global al final del intervalo dado por $EG(1) = e - y[N]$.

```
(%i4)  f(t,y):=y$ Taylor3(f,0,1,100,1)$
```

(%t5)

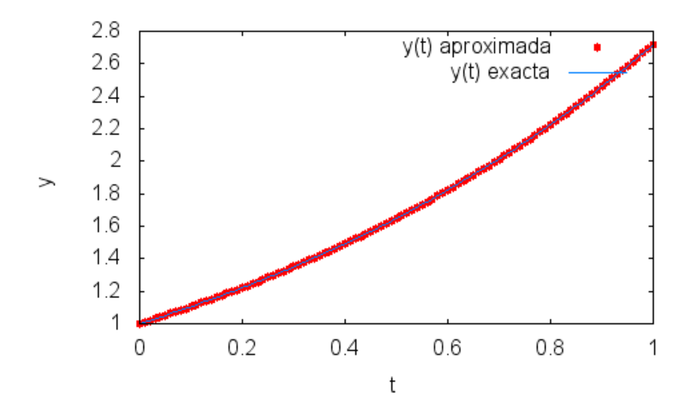


Figura 6.2: Exacta vs aproximada por Taylor de orden 3 y h = 0.01

(%o6) El error global en 1,0 es $EG(1{,}0) = 1{,}1235941110854242 \cdot 10^{-7}$

Observaciones.

- El error se comporta como era de esperar para un método de tercer orden, pues al dividir el paso por 10 el error se ha dividido aproximadamente por $1000 = 10^3$.
- Se puede generalizar a cualquier orden basta con definir y añadir los términos correspondientes a la función fi. En particular, cuando fi se reduce a f se tiene el método de Euler explícito.
- En las salidas, se puede omitir la solución exacta, pues en general no se conoce, o detenernos sólo en los valores en determinados puntos del intervalo, por ejemplo en el punto final.

6.9.2. Métodos Runge-Kutta explícitos para EDO's

Para fijar ideas nos vamos a centrar en los métodos Runge-Kutta explícitos de tres etapas. El programa constará, en general, de varios bloques, en el primero de ellos hemos de definir el PVI a integrar, es decir los valores iniciales, la longitud del intervalo de integración y la ecuación diferencial, en el segundo hemos de introducir el tablero de Butcher correspondiente al método Runge-Kutta explícito de tres etapas que vamos a utilizar, luego debemos programar el esquema del método y la salida deseada, que puede ser el valor aproximado de la solución en el punto final del intervalo o los valores aproximados en toda la red de puntos considerada y/o su representación gráfica. También podemos determinar antes de comenzar el orden del método mediante las ecuaciones de orden.

En cada paso el esquema es siempre el mismo para cualquier método Runge-Kutta: se calculan los valores auxiliares k_i (en este caso k_1, k_2, k_3) y a partir de ellos se actualiza la solución y_n. También hay que actualizar el valor temporal $t_n = t_{n-1} + h$. Para evitar que la memoria se desborde se suelen sobrescribir tanto t_n como y_n, sobre todo si sólo deseamos presentar el valor en el punto final del intervalo, aunque si el número N no es muy grande se pueden guardar todos y hacer su gráfica como antes. Recordemos que cada k_i $(i = 1, 2, 3)$ es de la forma:

$$k[i] : f(t + c[i] * h, y + h * sum(a[i, j] * k[j], j, 1, i - 1))$$

En tanto que $y[n] : y[n-1] + h * sum(b[i] * k[i], i, 1, 3)$. Para dar un método RK(3) explícito hay que dar los coeficientes $b[i]$ $(i = 1, 2, 3)$, los $c[2], c[3]$, así como los $a[2, 1], a[3, 1]$ y $a[3, 2]$.

Vamos a integrar nuevamente el problema anterior con el método RK explícito de tres etapas y orden 3 para el que $b[1] = 1/6, b[2] = 4/6, b[3] = 1/6, c[2] = 1/2, c[3] = 1, a[2,1] = 1/2, a[3,1] = -1$ y $a[3,2] = 2$, ahora como se cumplen las ecuaciones 1), 2) y 3) anteriores, el método en cuestión, denominado método de Kutta, es de tercer orden. Se puede hacer un programa Maxima mediante un block, en cuyos argumentos se define la ecuación a integrar, los valores iniciales, el número de subdivisiones del intervalo de integración y su amplitud, como el que sigue

```
(%i6)  RK3(f,t0,y0,N,T):= block(
       [numer], numer:true, kill(t,y),

       /*Tamaño del paso y valores iniciales*/
       h:T/N, t[0]:t0, y[0]:y0,

       /*Coeficientes del método*/
       b[1]:1/6, b[2]:4/6, b[3]:1/6, c[2]:1/2,
       c[3]:1, a[2,1]:1/2, a[3,1]:-1, a[3,2]:2,

       /*Programa del método*/
       define(k1(t,y),f(t,y)),
       define(k2(t,y),f(t+c[2]*h,y+a[2,1]*h*k1(t,y))),
       define(k3(t,y),f(t+c[3]*h,y+h*(a[3,1]*k1(t,y)
       +a[3,2]*k2(t,y)))),

       for k:1 thru N do (t[k]:t[0]+k*h,
       y[k]:y[k-1]+h*(b[1]*k1(t[k-1],y[k-1])+
       b[2]*k2(t[k-1],y[k-1])+b[3]*k3(t[k-1],y[k-1]))),

       /*Salida numérica y gráfica discreta*/
       sol:makelist([t[j],y[j]],j,0,N)),
       print("Solución numérica: ",sol),
       wxplot2d([[discrete,sol]],[t,0,1],[style,[points,2,2]],
       [legend,"y(t) aproximada"],[xlabel,"t"],
       [ylabel,"y"]), display(y[N])
       )$
```

Apliquémoslo, en primer lugar, al problema anterior con $N = 10, h = 0{,}1$, para ello, basta hacer lo que sigue.

```
(%i10) f(t,y):=y$ RK3(f,0,1,10,1)$
       print("El error global en 1.0 es EG(1.0) = ",
       float(%e-y[10]))$
```

Solución numérica: la omitimos, sólo mostramos el valor de y_{10}
(%t8)

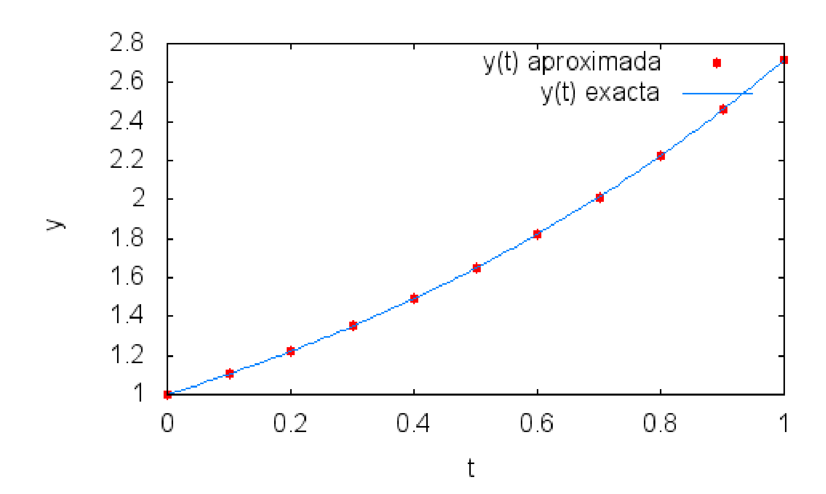


Figura 6.3: Exacta vs aproximada por RK(3) dado con h = 0.1

(%o9) $y_{10} = 2{,}718177262481609$

El error global en 1.0 es $EG(1{,}0) = 1{,}045659774359109 \cdot 10^{-4}$ que es similar al del método de Taylor del mismo orden. La programación de otros métodos Runge-Kuutta explícitos es similar.

6.9.3. El paquete diffeq y el comando rk

Maxima dispone de un paquete denominado "**diffeq**" que permite integrar numéricamente PVI para ecuaciones o sistemas de ecuaciones diferenciales de primer orden mediante el comando "**rk**". Lo primero de todo es cargar este paquete mediante el comando "**load(diffeq)**" y luego utilizar el comando "**rk(f, y, y0, [t, t0, t1, h])**", cuyos argumentos son la función f que define la ecuación o sistema diferencial ordinaria de primer orden, y es la variable función incógnita, $y0$ son los valores iniciales de y, $[t, t0, t1, h]$ son la variable independiente t y sus valores inicial y final, h el paso de integración, hay que llevar cuidado pues, a veces, debido a los errores de redondeo

sumando muchas veces h a $t0$ no llegamos exactamente a $t1$, y es necesario dar un paso final para ajustar el valor de la solución en el punto $t1$; la salida produce el valor aproximado de la solución en los puntos $t0 + ih$ siempre que estos sean menores que $t1$, las aproximaciones son obtenidas aplicando el método Runge-Kutta "clásico" de cuatro etapas y orden cuatro. Veamos su aplicación al PVI anterior con $h = 0{,}1$ y $h = 0{,}01$.

```
(%i10) kill(all)$ load(diffeq)$
       kill(f)$ f(t,y):=y$
       print("Solución numérica:")$
       numsol:rk(f(t,y),y,1,[t,0,1,0.1]);
       print("Representación gráfica de la solución numérica:")$
       wxplot2d([discrete,numsol],[style,points])$
```

Solución numérica:
(%o5) [[0,1],[0.1,1.105170833333333],[0.2,1.221402570850695],
[0.3,1.349858497062538],[0.4,1.491824240080686],[0.5,1.648720638596838],
[0.6,1.822117962091933],[0.7,2.013751626596777],[0.8,2.225539563292315],
[0.9,2.459601413780071],[1.0,2.718279744135166]]
Representación gráfica de la solución numérica:
(%t7)

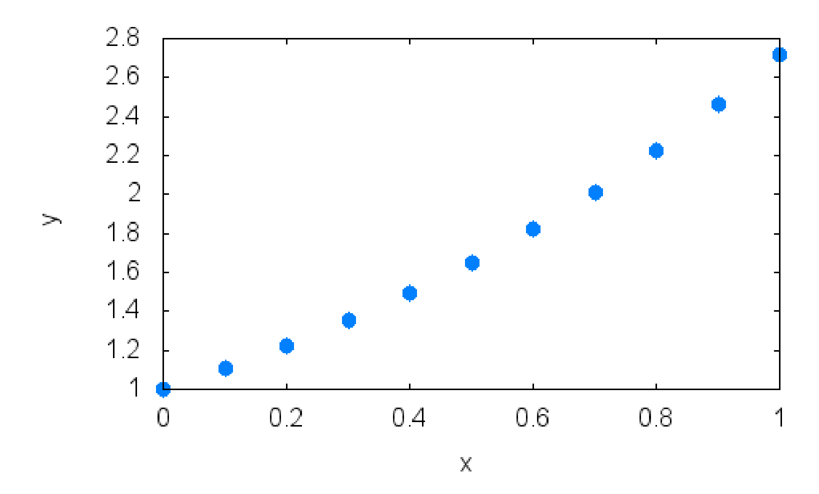


Figura 6.4: Solución aproximada por RK(4) "clásico" con h = 0.1

```
(%i8) print("El error global en 1.0 es EG(1.0) = ",
      float(%e-2.718279744135166))$kill(f)$
```

El error global en 1.0 es $EG(1{,}0) = 2{,}0843238792700447 \cdot 10^{-6}$

```
(%i10) f(t,y):= y;
       solnum:rk(f(t,y),y,1.0,[t,0.0,1.0,0.01]);
       wxplot2d([discrete,solnum],[style,points])$
```

(%o10) $\mathrm{f}\,(t, y) := y$
(%o11) Omitimos la salida numérica para abreviar.
(%t12)

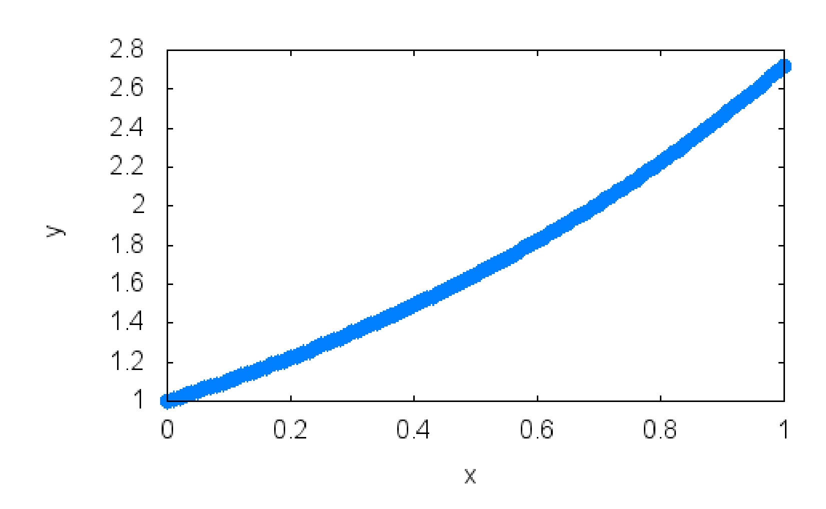


Figura 6.5: Solución aproximada por RK(4) “clásico” con h = 0.01

```
(%i13) print("El error global en 1.0 es EG(1.0) = ",
       float(%e-2.718281828234403))$
```

El error global en 1.0 es $EG(1{,}0) = 2{,}2464208271344432 \cdot 10^{-10}$

Aplicación a la integración numérica del problema plano de dos cuerpos

Consideremos el problema plano de dos cuerpos puntuales, que se mueven atraídos según la ley de la gravitación universal de Newton. En unidades apropiadas, las ecuaciones del movimiento adoptan la forma:

$y1' = y3$
$y2' = y4$
$y3' = -y1/(y1^2 + y2^2)^{3/2}$

$y4' = -y2/(y1^2 + y2^2)^{3/2}$
donde $(y1, y2)$ son las coordenadas de uno de los cuerpos (satélite) en un sistema con origen en el otro cuerpo central, en tanto que $(y3, y4)$ representa el vector velocidad del cuerpo satélite.

Como recordaréis por la primera ley de Kepler: "los planetas describen órbitas elípticas alrededor del sol, que ocupa uno de los focos de esa elipse". En general, las soluciones del problema de dos cuerpos son cónicas (pueden ser elipses, parábolas o hipérbolas), pero si se toman las condiciones iniciales siguientes: $y1(0) = 1, y2(0) = y3(0) = 0, y4(0) = 1$), la solución es una circunferencia de centro el origen radio 1 y con periodo 2π. Con ayuda del editor de Maxima este PVI puede escribirse en la forma:

```
(%i14) kill(all)$
       'diff(y1,t) = y3;
       'diff(y2,t) = y4;
       'diff(y3,t) = -y1/(y1^2+y2^2)^(3/2);
       'diff(y4,t) = -y2/(y1^2+y2^2)^(3/2);
       print("y(0) = ", y:[1.0,0.0,0.0,1.0])$
```

(%o1) $\frac{d}{dt} y1 = y3$

(%o2) $\frac{d}{dt} y2 = y4$

(%o3) $\frac{d}{dt} y3 = -\frac{y1}{\left(y2^2 + y1^2\right)^{\frac{3}{2}}}$

(%o4) $\frac{d}{dt} y4 = -\frac{y2}{\left(y2^2 + y1^2\right)^{\frac{3}{2}}}$

$y(0) = [1{,}0, 0{,}0, 0{,}0, 1{,}0]$

El esquema es similar al de antes, pero ahora es una función vectorial de cuatro componentes y es ligeramente más complicado, por ello vamos a recurrir al comando de Maxima rk para resolver este sistema, puesto que la solución es periódica con periodo 2π, si integramos de 0 a 2π, debemos tener valores próximos a los iniciales, veamos ahora la solución obtenida con el paso $h = 2\pi/20 = \pi/10$.

```
(%i6) /* Con 20 pasos */
      kill(all)$ load(diffeq);numer:true$ kill(f)$
      y:[y1,y2,y3,y4];h=%pi/10;
      f(t,y):=[y3,y4,-y1/(y1^2+y2^2)^(3/2),
      -y2/(y1^2+y2^2)^(3/2)];
      discretesolution1:rk(f(t,y),y,[1.0,0.0,0.0,1.0],
      [t,0,2*%pi,%pi/10])$ last(discretesolution1);
```

(%o1) $C : /PROGRA\ 2/MAXIMA\ 1{,}0-2/share/maxima/5{,}28{,}0-2/share/numeric/diffeq.mac$

(%o4) $[y1, y2, y3, y4]$

(%o5) $h = 0{,}31415926535898$

(%o6) $\mathrm{f}\,(t, y) := [y3, y4, \frac{-y1}{(y1^2 + y2^2)^{\frac{3}{2}}}, \frac{-y2}{(y1^2 + y2^2)^{\frac{3}{2}}}]$

(%o8) [6.283185307179586,0.99944519714206,0.0038657399331735, -0.0038695061807764,1.000266720696481]

La salida (%o8) anterior, correspondiente al comando last(discretesolution1), es el valor correspondiente a $t = 20h = 6{,}283185307179586 \leq 2\pi$, que lo calcula, puesto que en los redondeos resulta $20h \leq 2\pi$, y es la solución aproximada para $t = 2\pi = 6{,}283185307179586$, siendo el error global:

```
(%i9) print("El error global en 2*%pi es EG(",2*%pi,") = ",
      [1.0,0.0,0.0,1.0]-[0.99944519714206,0.0038657399331735,
      -0.0038695061807764,1.000266720696481])$
```

El error global en 2π es $EG(6{,}283185307179586) =$ $[5{,}548028579399622 \cdot 10^{-4}, -0{,}0038657399331735, 0{,}0038695061807764, -2{,}6672069648103758 \cdot 10^{-4}]$.

Este se ha obtenido aplicando el método con 20 pasos, ahora repitamos el proceso con 200 pasos.

```
(%i10) /*Con 200 pasos */ kill(all)$ load(diffeq);
       numer:true$ kill(f)$
       y:[y1,y2,y3,y4];h=%pi/100;
       f(t,y):=[y3,y4,-y1/(y1^2+y2^2)^(3/2),
       -y2/(y1^2+y2^2)^(3/2)];
       discretesolution2:rk(f(t,y),y,[1.0,0.0,0.0,1.0],
       [t,0,2*%pi,%pi/100])$ last(discretesolution2);
```

(%o1) $C : /PROGRA\ 2/MAXIMA\ 1{,}0-2/share/maxima/5{,}28{,}0-2/share/numeric/diffeq.mac$

(%o4) $[y1, y2, y3, y4]$

(%o5) $h = 0{,}031415926535898$

(%o6) $\mathrm{f}(t, y) := [y3, y4, \frac{-y1}{(y1^2 + y2^2)^{\frac{3}{2}}}, \frac{-y2}{(y1^2 + y2^2)^{\frac{3}{2}}}]$

(%o8) [6.251769380643689,0.9995065602185,-0.031410593663266, 0.03141059439279,0.99950656822774]

Observación. Como se puede observar, la última salida (%o8) da un error similar o peor que la anterior, a pesar de haber dividido el paso por 10. Se aprecia que el tiempo final para el que se calcula la aproximación es 6.251769380643689 distinto de 2π, lo cual se debe a los errores de redondeo, pues ahora se tiene que $6{,}251769380643689 + h > 2\pi$, por lo que no calcula el valor aproximado de la solución para $t = 2\pi$, como podemos comprobar a continuación.

```
(%i9)  is(6.251769380643689+0.031415926535898>2*%pi);
```

(%o9) *true*

Es pues necesario un ajuste del último paso, para obtener una mejor aproximación de la solución en 2π, haciendo un sólo paso de amplitud $2\pi - 6{,}251769380643689$ a partir del punto final anterior, en la forma:

```
(%i10) kill(all)$ load(diffeq);
       numer:true$ kill(f)$
       y:[y1,y2,y3,y4];
       f(t,y):=[y3,y4,-y1/(y1^2+y2^2)^(3/2),
       -y2/(y1^2+y2^2)^(3/2)];
       ajustesolution:rk(f(t,y),y,[0.9995065602185,
       -0.031410593663266,0.03141059439279,0.99950656822774],
       [t,6.251769380643689,2*%pi,2*%pi-6.251769380643689]);
```

(%o1) $C : /PROGRA\ 2/MAXIMA\ 1{,}0 - 2/share/maxima/5{,}28{,}0 - 2/share/numeric/diffeq.mac$

(%o4) $[y1, y2, y3, y4]$

(%o5) $\mathrm{f}(t, y) := [y3, y4, \frac{-y1}{(y1^2 + y2^2)^{\frac{3}{2}}}, \frac{-y2}{(y1^2 + y2^2)^{\frac{3}{2}}}]$

(%o6) $[[6{,}251769380643689, 0{,}9995065602185, -0{,}031410593663266, 0{,}03141059439279, 0{,}99950656822774], [6{,}283185307179586, 0{,}99999999465787, 1{,}6532531159352271\,10^{-7}, -1{,}6532530891510966\,10^{-7}, 1{,}000000002671051]]$

Con lo cual el error global en 2π estará dado ahora por el comando

```
(%i7) print("El error global en 2*%pi es EG(",2*%pi,") = ",
      [1.0,0.0,0.0,1.0]- [0.99999999465787,
      1.6532531159352271*10^-7,-1.6532530891510966*10^-7,
      1.000000002671051])$
```

El error global en 2π es $EG(6{,}283185307179586) =$
$[5{,}3421299606171146\,10^{-9}, -1{,}6532531159352271\,10^{-7}, 1{,}6532530891510966\,10^{-7}, -2{,}6710509359872958\,10^{-9}]$

6.10. Problemas y trabajos propuestos

Problemas propuestos:

1. Usar el método de Euler y la extrapolación de Richardson para calcular una aproximación de $y(0{,}2)$, siendo $y(t)$ la solución del problema de valor inicial $y' = t^2 - y;\ y(0) = 2$, tomar $h = 0{,}2$ y $h = 0{,}1$ y dar una estimación del error cometido en la aproximación obtenida con paso $h = 0{,}1$.

2. El problema de valor inicial $y' = \sqrt{y};\ y(0) = 0$, tiene la solución no trivial $y(t) = (\frac{t}{2})^2$, pero si aplicamos el método de Euler, obtenemos la solución $\nu(t, h) = 0$ para todo t y $h = \frac{t}{n}$ $(n = 1, 2, \cdots)$, explicar esta paradoja.

3. Resuelve numéricamente el problema de valor inicial $y' = 1-y;\ y(0) = 0$, por el método de Taylor de orden dos, utilizando los pasos $h = 0{,}5$, $h = 0{,}25$ y $h = 0{,}125$; comparar los valores obtenidos de la solución numérica en$t_n = 1$ con los valores de la solución exacta y estimar el error cometido. Asimismo, mejorar los datos obtenidos por el método de extrapolación al límite de Richardson.

4. Probar que el método de un paso asociado a la función $\Phi(t, y, h)$ definida por
$$\Phi(t, y, h) = f(t, y) + \frac{h}{2} f^{(1)}(t + \frac{h}{3}, y\frac{h}{3} f(t, y))$$
es de tercer orden.

5. Probar que para el problema de valor inicial $y' = ay;\ y(0) = 1$, el método Runge-Kutta clásico de cuarto orden da el mismo algoritmo que el método de Taylor de cuarto orden.

6. Estudiar la convergencia y el orden del método multipaso

$$y_{n+2} - y_n = 2hf_{n+1}$$

   Este método es conocido como la regla del punto medio.

7. ¿Cuál de los siguientes métodos lineales multipaso es convergente?

   a) $y_n - y_{n-2} = h(f_n - 3f_{n-1} + 4f_{n-2})$
   b) $y_n - 2y_{n-1} + y_{n-2} = h(f_n - f_{n-1})$
   c) $y_n - y_{n-1} - y_{n-2} = h(f_n - f_{n-1})$
   d) $y_n - 3y_{n-1} + 2y_{n-2} = h(f_n + f_{n-1})$
   e) $y_n - y_{n-2} = h(f_n - 3f_{n-1} + 2f_{n-2})$

Trabajos propuestos:

Se propone en este tema realizar alguno de los siguientes trabajos:

- Series de Butcher y estudio del orden de los métodos Runge-Kutta.
- Métodos Runge-Kutta implícitos.
- Métodos Runge-Kutta encajados: pares de Fehlberg.
- Estabilidad absoluta lineal de los métodos Runge-Kutta.
- Estabilidad absoluta lineal de los métodos multipaso.

Capítulo 7

Métodos en diferencias finitas para problemas de contorno

7.1. Introducción

Hasta ahora hemos estudiado la solución numérica de PVI, pero hay otro tipo de problemas cuya solución interesa en la práctica, como por ejemplo, hallar la solución de

$$PC\left\{\begin{array}{l} y''(t) = f(t, y(t), y'(t)), \forall t \in [a, b] \\ y(a) = \alpha, \quad y(b) = \beta \end{array}\right. \tag{7.1}$$

que es un problema de valor en la frontera (o de contorno) en dos puntos. Tales problemas son, en general, más difíciles de resolver que los problemas de valor inicial. Consideremos un ejemplo sencillo

$$\left\{\begin{array}{l} y''(t) = -y(t)t \in [0, \pi/2] \\ y(0) = 3, \quad y(\pi/2) = 7 \end{array}\right.$$

su solución general sería $y(t) = Asent + Bcost$, e imponiendo las condiciones $y(0) = 3$ e $y(\pi/2) = 7$, tenemos $A = 7$ y $B = 3$, luego la solución sería $y(t) = 7sent + 3cost$. Pero, este modo de resolución no es, en general, viable como puede verse con unos sencillos ejemplos.

Antes de considerar los métodos numéricos utilizados para resolver este tipo de problemas, veamos algunos resultados relativos a la existencia y unicidad de soluciones en este tipo de problemas; resultados poco generales y complicados de establecer.

Teorema 51 *Sea el problema de contorno*

$$PC\left\{\begin{array}{l} y''(t) = f(t, y(t), y'(t)), \forall t \in [a, b] \\ y(a) = \alpha, \quad y(b) = \beta \end{array}\right.$$

donde $f : D = [a, b] \times \mathbb{R}^2 \to \mathbb{R}$*, con* a *y* b *reales finitos, es tal que* $f(t, y, y')$*,* $\frac{\partial f(t,y,y')}{\partial y}$ *y* $\frac{\partial f(t,y,y')}{\partial y'}$ *son continuas en* D*. Entonces, si se verifican las condiciones:*

1. $\frac{\partial f(t,y,y')}{\partial y} > 0 \ \forall (t, y, y') \in D$
2. $\exists \mu$ *tal que* $\mid \frac{\partial f(t,y,y')}{\partial y'} \mid \leq \mu, \ \forall (t, y, y') \in D$

el problema tiene una única solución.

Observación: Cuando f puede ser expresada en la forma $f(t, y, y') = p(t)y' + q(t)y + r(t)$, la ecuación diferencial se dice lineal y el teorema anterior se simplifica en el siguiente.

Corolario 3 *Dado el problema de contorno*

$$PC \left\{ \begin{array}{l} y''(t) = p(t)y' + q(t)y + r(t), \forall t \in [a, b] \\ y(a) = \alpha, \quad y(b) = \beta \end{array} \right. \tag{7.2}$$

verificando las condiciones

1. $p(t)$*,* $q(t)$ *y* $r(t)$ *son continuas en* $[a, b]$*.*
2. $q(t) > 0, \ \forall t \in [a, b]$

el problema tiene solución única.

Ejercicio.- Probar que el problema de contorno:

$$\left\{ \begin{array}{l} y'' + e^{-ty} + seny' = 0, \forall t \in [1, 2] \\ y(1) = y(2) = 0 \end{array} \right.$$

tiene solución única (basta aplicar el teorema anterior).

7.2. Métodos en diferencias finitas para problemas de contorno lineales

Los métodos numéricos más utilizados en la práctica para el cálculo de soluciones aproximadas de este problema son los métodos de tiro y de diferencias finitas, pasamos a describir este último. La idea básica de este tipo de métodos consiste en la aproximación de las derivadas en las ecuaciones diferenciales por cocientes en diferencias, adecuadamente elegidos para mantener un cierto orden del error de truncamiento. Este tipo de métodos es

preferido con relación a los anteriores debido a los problemas de inestabilidad, si bien requieren un poco más de computación para una aproximación deseada. Veamos el caso de un problema de contorno lineal.

Sea el problema de contorno lineal (7.2) escrito en la forma:

$$PC \begin{cases} -y''(t) + p(t)y' + q(t)y + r(t) = 0, \forall t \in [a, b] \\ y(a) = \alpha, \quad y(b) = \beta \end{cases}$$

para su resolución numérica, elegimos una malla de puntos equidistantes $t_j = a + jh, \; j = 0, 1, 2, \ldots, n + 1$, con un tamaño de paso h dado por $h = \frac{b-a}{n+1}$, en los puntos interiores de la malla $\{t_j, \; j = 1, 2, \ldots, n\}$ podemos utilizar las siguientes aproximaciones de las derivadas:

$$\begin{cases} y''(t_j) \cong \frac{y(t_{j+1})-2y(t_j)+y(t_{j-1})}{h^2} \\ y'(t_j) \cong \frac{y(t_{j+1})-y(t_{j-1})}{2h} \end{cases} \tag{7.3}$$

con errores de truncamiento de la forma $O(h^2)$, obtenidos en el capítulo 5 anterior, suponiendo que $y(t)$ es suficientemente diferenciable. Para obtener el sistema de ecuaciones, asociado al problema de contorno (7.2) dado, tomemos $y_0 = \alpha, y_{n+1} = \beta$ y llamemos $y_1, y_2, \ldots, y_n$ a las aproximaciones buscadas de la solución en los puntos interiores $\{t_j, \; j = 1, 2, \ldots, n\}$, al sustituir (7.3) en la ecuación se obtienen las ecuaciones en diferencias:

$$\frac{-y_{j-1} + 2y_j - y_{j+1}}{h^2} + p(t_j)\frac{y_{j+1} - y_{j-1}}{2h} + q(t_j)y_j + r(t_j) = 0, \; (j = 1, 2, \ldots, n)$$

tras multiplicar por h^2 y ordenar dichas ecuaciones, resulta que y_1, y_2,..., y_n deben verificar el sistema lineal:

$$-(1+\frac{h}{2}p(t_j))y_{j-1}+(2+h^2q(t_j))y_j-(1-\frac{h}{2}p(t_j))y_{j+1} = -h^2r(t_j), \; (j = 1, 2, \ldots, n) \tag{7.4}$$

que puede expresarse en la forma abreviada

$$Ay = b$$

siendo A, b e y las matrices siguientes:

$$A = \begin{bmatrix} 2 + h^2q(t_1) & -1 + \frac{h}{2}p(t_1) & 0 & \ldots & 0 \\ -1 - \frac{h}{2}p(t_2) & 2 + h^2q(t_2) & -1 + \frac{h}{2}p(t_2) & \ldots & 0 \\ 0 & -1 - \frac{h}{2}p(t_3) & 2 + h^2q(t_3) & \ldots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & -1 - \frac{h}{2}p(t_{n-1}) & 2 + h^2q(t_{n-1}) \end{bmatrix}$$

$$b = \begin{bmatrix} -h^2 r(t_1) + (1 + \frac{h}{2} p(t_1)) y_0 \\ -h^2 r(t_2) \\ \vdots \\ -h^2 r(t_{n-1}) \\ -h^2 r(t_n) + (1 - \frac{h}{2} p(t_n)) y_{n+1} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}$$

Y para el que se tiene el siguiente resultado.

Teorema 52 *Supongamos que $p(t), q(t), r(t)$ son continuas en $[a, b]$. Entonces, si $q(t) \geq 0$ sobre $[a, b]$, el sistema lineal anterior tiene una única solución, siempre que $h < 2/L$, siendo $L = \max_{t \in [a,b]} \mid p(t) \mid$.*

Observaciones:

1. Si la solución del problema de contorno lineal es de clase $C^{(4)}([a, b])$, entonces, el error de la aproximación en diferencias finitas es del orden de $O(h^2)$. Puede mejorarse por el método de extrapolación al límite de Richardson considerando una reducción en el tamaño del paso, pues el error se puede expresar mediante un desarrollo de potencias pares de h con coeficientes independientes del paso h (ver Keller (1968)).

2. Para el caso no lineal, el método es similar, si bien hay que resolver un sistema de n ecuaciones no lineales, para lo cual podemos utilizar el método de Newton para sistemas no lineales que requiere que el valor inicial esté suficientemente próximo a la raíz buscada y que la matriz Jacobiana correspondiente sea inversible.

7.3. Generalidades sobre ecuaciones en derivadas parciales

Vamos a considerar tan sólo ecuaciones en derivadas parciales lineales de segundo orden, a las que pertenecen las denominadas ecuaciones de la Física Matemática. Si el número de variables independientes es dos, denominándolas por x e y, su forma general es

$$a(x,y)\frac{\partial^2 u}{\partial x^2}+b(x,y)\frac{\partial^2 u}{\partial x \partial y}+c(x,y)\frac{\partial^2 u}{\partial y^2}+d(x,y)\frac{\partial u}{\partial x}+e(x,y)\frac{\partial u}{\partial y}+f(x,y)u = g(x,y) \tag{7.5}$$

donde u es una función a determinar sobre algún dominio Ω, verificando la ecuación (7.5) y algunas condiciones iniciales y/o de contorno sobre la frontera. Si $g(x, y) = 0$ la ecuación se dice homogénea y en caso contrario completa o no homogénea. Atendiendo a los términos conteniendo las derivadas de mayor orden, la ecuación (7.5) se dice que en el punto (x, y) es del tipo:

1. Hiperbólico si $b^2 - 4ac > 0$
2. Parabólico si $b^2 - 4ac = 0$
3. Elíptico si $b^2 - 4ac < 0$

Si $a(x, y)$, $b(x, y)$ y $c(x, y)$ son constantes en la región Ω, el tipo también lo es. Ejemplos paradigmáticos de cada una de ellas son las ecuaciones siguientes:

1. $u_{tt} = c^2 u_{xx}$ es la ecuación de ondas unidimensional (hiperbólica)
2. $u_t = \kappa u_{xx}$ es la ecuación del calor unidimensional (parabólica)
3. $u_{xx} + u_{yy} = 0$ es la ecuación de Laplace bidimensional (elíptica)

Para la resolución numérica de las ecuaciones anteriores utilizaremos los denominados métodos en diferencias finitas, en los que las derivadas de una función se sustituyen por cocientes de diferencias; además, se considerará una malla de puntos (x_i, y_j) para el caso elíptico y (x_i, t_j) para los casos hiperbólico o parabólico, definidos por

$$\begin{cases} x_i = x_0 + ih, & h \textit{ constante positiva} \\ y_j = y_0 + jk, & (o\ t_j = t_0 + jk)\ k \textit{ constante positiva} \end{cases}$$

Los valores de cualquier función en dichos puntos se denotarán por $h_{ij} = h(x_i, y_j)$ (o bien $h_{ij} = h(x_i, t_j)$), en tanto que las aproximaciones o discretizaciones por diferencias finitas y sus errores de truncatura más frecuentes para las derivadas de u en tales puntos, con hipótesis adecuadas de diferenciabilidad, adoptan la forma:

Aproximaciones	Errores de truncatura
$u_x(x_i, t_j) \cong \frac{u_{i+1,j} - u_{i,j}}{h}$	$O(h)$
$u_t(x_i, t_j) \cong \frac{u_{i,j+1} - u_{i,j}}{k}$	$O(k)$
$u_x(x_i, t_j) \cong \frac{u_{i,j} - u_{i-1,j}}{h}$	$O(h)$
$u_t(x_i, t_j) \cong \frac{u_{i,j} - u_{i,j-1}}{k}$	$O(k)$
$u_x(x_i, t_j) \cong \frac{u_{i+1,j} - u_{i-1,j}}{2h}$	$O(h^2)$
$u_t(x_i, t_j) \cong \frac{u_{i,j+1} - u_{i,j-1}}{2k}$	$O(k^2)$
$u_{xx}(x_i, t_j) \cong \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}$	$O(h^2)$
$u_{yy}(x_i, y_j) \cong \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2}$	$O(k^2)$
$u_{xy}(x_i, y_j) \cong \frac{u_{i+1,j+1} + u_{i-1,j-1} - u_{i+1,j-1} - u_{i-1,j+1}}{4hk}$	$O(h^2 + k^2)$

Ejercicio.- Con condiciones adecuadas de diferenciabilidad sobre la función u, obtener las expresiones explícitas de los errores de truncatura de la tabla anterior.

7.4. Ecuaciones elípticas: Problemas de valor en la frontera para ecuaciones de Poisson o Laplace

Una ecuación de **Poisson** con condiciones de contorno tipo **Dirichlet** es una ecuación del tipo

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f \tag{7.6}$$

con las condiciones de contorno

$$u(x,y) = g(x,y),\ \forall (x,y) \in \Gamma = \partial\Omega\ (frontera\ de\ \Omega) \tag{7.7}$$

Representa problemas independientes del tiempo. Un caso particular interesante del anterior es aquel en el que $f \equiv 0$, se denomina ecuación de **Laplace**, así el problema:

$$\begin{cases} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \\ u(x,y) = g(x,y), \forall (x,y) \in \Gamma \end{cases} \tag{7.8}$$

Representaría, por ejemplo una distribución de temperatura en Ω en estado estacionario. A veces nos dan la variación normal de u en la frontera Γ de Ω, tenemos entonces una condición de contorno de tipo **Newmann**

$$\frac{\partial u}{\partial n} = g,\ en\ \Gamma$$

o bien, una condición mixta

$$\alpha \frac{\partial u}{\partial n} + \beta u = g\ en\ \Gamma$$

Consideremos el problema de Poisson bidimensional (7.6) en el rectángulo $\Omega = [0,a] \times [0,b]$, dividamos el intervalo $[0,a]$ en $n+1$ subintervalos iguales de longitud $h = \frac{a}{n+1}$ y el $[0,b]$, en $m+1$ subintervalos iguales de longitud $k = \frac{b}{m+1}$, con ello se forma una malla bidimensional, cuyos nodos (x_i, y_j) son $x_i = ih$ e $y_j = jk$; llamaremos $H = (h,k)$ y $|H| = \text{máx}\{h,k\}$ (norma de la partición). Sean los conjuntos discretos

$$\begin{array}{l} \Omega_H = \{(x_i, y_j) | 1 \le i \le n,\ 1 \le j \le m\} \\ \Gamma_H = \{(x_0, y_0), \ldots, (x_{n+1}, y_0), (x_0, y_{m+1}), \ldots, (x_n, y_{m+1})\} \end{array} \tag{7.9}$$

constituido el primero por los nm puntos de la malla interiores a Ω y el segundo por los $2(n+m)+4$ puntos de la malla pertenecientes a la frontera Γ de Ω. Entonces, si u es de clase $C^{(4)}$ en un entorno de Ω, tras sustituir u_{xx}

y u_{yy} en cada uno de los nm puntos interiores (x_i, y_j) por sus aproximaciones de segundo orden, se obtienen las ecuaciones

$$\boxed{\begin{array}{l}\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} + \frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{k^2} = f_{i,j} \\ (1 \leq i \leq n, 1 \leq j \leq m)\end{array}} \tag{7.10}$$

que constituyen un sistema lineal de nm ecuaciones y $nm+2(n+m)$ incógnitas, pues por la forma de discretizar no figuran los valores de u en los vértices del rectángulo, ahora bien por las condiciones de contorno

$$u_{i,j} = g_{i,j} \; en \; \Gamma_H \tag{7.11}$$

son conocidos los valores de u en los $2(n+m)$ puntos de la frontera que aparecen en las mismas, por tanto (7.10) se reduce a un sistema de nm ecuaciones lineales de coeficientes constantes con el mismo número de incógnitas $u_{i,j}$, que serán las aproximaciones de la solución buscada en los puntos interiores de la malla. Para la resolución de este sistema si el número de ecuaciones es alto, que es lo normal, se suelen utilizar métodos iterativos tipo Jacobi, Gauss-Seidel o relajación.

En relación a este problema, podemos enunciar los dos resultados siguientes.

Proposición 7 *El problema discreto planteado tiene solución única.*

Proposición 8 *El error de truncatura local del problema puede escribirse como sigue*

$$\tau(u) = -[\frac{h^2}{12}\frac{\partial^4 u(x_i + \theta_i h, y_j)}{\partial x^4} + \frac{k^2}{12}\frac{\partial^4 u(x_i, y_j + \theta_j k)}{\partial y^4}] \tag{7.12}$$

y si las derivadas cuartas anteriores de la solución u están acotadas por μ_1 y μ_2 respectivamente en Ω_H se tiene

$$\mid \tau(u) \mid \leq \frac{1}{12}[h^2\mu_1 + k^2\mu_2] \tag{7.13}$$

Y si la norma de la partición tiende a cero se tiene la convergencia, es decir $u_{i,j} \longrightarrow u(x_i, y_j)$.

7.5. Ecuaciones parabólicas: la ecuación del calor

Consideraremos en este apartado diversos esquemas explícitos e implícitos para la resolución numérica de la ecuación parabólica unidimensional, escrita

en forma canónica como $u_t = u_{xx}$ sujeta a condiciones de contorno e iniciales que garantizan la existencia y unicidad de soluciones del problema, dicha ecuación se suele utilizar para modelizar la distribución de temperaturas $u(x_i, t_j)$ en los puntos x_i de una varilla finita en cada instante t_j.

Método explícito en diferencias finitas. Sin pérdida de generalidad, consideraremos el problema de difusión del calor en una varilla de longitud unidad con $\kappa = 1$, dado por la ecuación

$$u_t = u_{xx}, \ 0 < x < 1, \ t \in \mathbb{R}^+ \tag{7.14}$$

con las condiciones de frontera

$$u(0,t) = u(1,t) = 0, \ t \in \mathbb{R}^+ \tag{7.15}$$

y la condición inicial

$$u(x,0) = f(x), \ 0 < x < 1 \tag{7.16}$$

Elijamos dos tamaños de paso constantes h, k, de modo que $1/h = n$, sea un número entero y k sea un número real positivo arbitrario. Así, tendremos definidos los nodos espaciales $x_i = ih$ $(i = 0, 1, \ldots, n)$ y los tiempos discretos $t_j = jk$ $(j = 0, 1, \ldots, m)$. De modo que la solución numérica vendrá dada por un conjunto de valores u_{ij}, que serán las aproximaciones de la solución exacta en los puntos: (x_i, t_j), es decir u_{ij} es una aproximación a la temperatura del punto de la varilla de abscisa x_i en el tiempo t_j. En condiciones adecuadas (por ejemplo si u es al menos de clase $\mathcal{C}^{(4)}$ en su dominio de definición) la fórmula de Taylor nos permite escribir el desarrollo

$$u(x_i, t_j+k) = u(x_i, t_j)+ku_t(x_i, t_j)+\frac{k^2}{2}u_{tt}(x_i, t_j+\theta_j k) \ \mathit{con} \ 0 < \theta_j < 1 \tag{7.17}$$

por tanto se tiene

$$u_t(x_i, t_j) = \frac{u(x_i, t_j + k) - u(x_i, t_j)}{k} - \frac{k}{2}u_{tt}(x_i, t_j+\theta_j k) \ \mathit{con} \ 0 < \theta_j < 1 \tag{7.18}$$

estando dada la aproximación de la derivada temporal en diferencias progresivas por

$$u_t(x_i, t_j) \cong \frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} \tag{7.19}$$

Análogamente, los desarrollos de Taylor respecto a la variable x proporcionan la fórmula

$$\begin{array}{l} u_{xx}(x_i, t_j) = \frac{u(x_i+h, t_j)-2u(x_i, t_j)+u(x_i-h, t_j)}{h^2} - \frac{h^2}{12}u_{xxxx}(x_i + \xi_i h, t_j) \\ \mathit{con} \quad -1 < \xi_i < 1 \end{array} \tag{7.20}$$

de donde deducimos la siguiente aproximación en diferencias centrales para las derivadas espaciales

$$u_{xx}(x_i,t_j) \cong \frac{u(x_{i+1},t_j) - 2u(x_i,t_j) + u(x_{i-1},t_j)}{h^2} \tag{7.21}$$

Utilizando la aproximaciones de las derivadas (7.19) y (7.21), y representando por u_{ij}, el valor aproximado de $u(x_i,t_j)$, nos quedan las ecuaciones en diferencias

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} \tag{7.22}$$

En tanto que, el error de truncamiento local viene dado por la fórmula

$$\tau_{ij} = \frac{k}{2} u_{tt}(x_i, t_j + \theta_j k) - \frac{h^2}{12} u_{xxxx}(x_i + \xi_i h, t_j) \tag{7.23}$$

De la ecuación aproximada (7.22) puede deducirse el algoritmo

$$\boxed{\begin{array}{l} u_{i,j+1} = u_{i,j}(1 - \frac{2k}{h^2}) + \frac{k}{h^2}(u_{i+1,j} + u_{i-1,j}) \\ (i = 1, \ldots, n-1;\ j = 0, 1, \ldots, m) \end{array}} \tag{7.24}$$

Puesto que $u(x,0) = f(x)$ en $0 < x < 1$, se puede iniciar el cálculo en (7.24) computando $u_{i1}, u_{i2}, \ldots, u_{i,m-1}$. Concretamente, de la condición inicial $u(x,0) = f(x)$ en $0 < x < 1$, resultan los valores

$$u_{i,0} = f(x_i),\ i = 1, 2, \ldots, n-1$$

Por otro lado, de las condiciones de frontera $u(0,t) = u(1,t) = 0,\ \forall t \in \mathbb{R}^+$ resultan

$$u_{0j} = u_{nj} = 0,\ j = 1, 2, \ldots, m$$

Utilizando (7.24) se calculan los $u_{i,j}$, procediéndose así m veces hasta el instante de tiempo $t = t_m$ deseado.

Definiendo el término $r = \frac{k}{h^2}$, que se conoce como **número o constante de Courant**, el algoritmo (7.24) se puede escribir en forma matricial como sigue

$$\boxed{U_j = AU_{j-1} \quad (j = 1, 2, \ldots, m)} \tag{7.25}$$

donde

$$A = \begin{bmatrix} 1-2r & r & \ldots & 0 & 0 \\ r & 1-2r & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1-2r & r \\ 0 & 0 & \ldots & r & 1-2r \end{bmatrix}$$

$$U_j = \begin{bmatrix} u_{1,j} \\ \vdots \\ \vdots \\ \vdots \\ u_{n-1,j} \end{bmatrix} \quad U_0 = \begin{bmatrix} f(x_1) \\ \vdots \\ \vdots \\ \vdots \\ f(x_{n-1}) \end{bmatrix}$$

El esquema en diferencias considerado se dice que es un método explícito en diferencias progresivas. Si la solución exacta $u(x,t)$ posee derivadas parciales continuas de cuarto orden, el método es de orden $O(k+h^2)$. El método se puede recordar por la denominada "**molécula computacional**" que indica que los nuevos valores $u_{i,j+1}$, se calculan explícitamente en términos de los previos $u_{i-1,j}, u_{i,j}, u_{i+1,j}$.

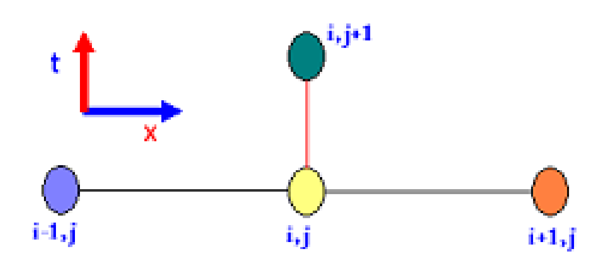


Figura 7.1: Molécula computacional del método explícito

Estudio de la estabilidad del método. El método que nos ocupa, se puede escribir en la forma matricial dada en (7.25), donde A es una matriz cuadrada de orden $(n-1)$, entonces, se puede deducir fácilmente que

$$U_j = A^j U_0, \ j = 1, 2, \dots \tag{7.26}$$

Ahora bien, del conocimiento de la solución explícita de este problema, que puede ser obtenida por separación de variables, o del hecho físico de que la temperatura en la barra tenderá a cero conforme $t \to \infty$, exigimos a la solución numérica que verifique la condición

$$\lim_{j \to \infty} A^j U_0 = 0, \ \forall U_0$$

lo cual ocurrirá sí y sólo si $\rho(A) < 1$. Y puesto que la matriz de iteración

$A = I - rB$, siendo I la matriz identidad de orden $n-1$ y B la matriz

$$B = \begin{bmatrix} 2 & -1 & \dots & 0 & 0 \\ -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & \dots & -1 & 2 \end{bmatrix}$$

todo valor propio de A es de la forma $\lambda_i = 1 - 2r\mu_i$, siendo μ_i un valor propio de B. Partiendo de esto, puede probarse que una **condición necesaria para la estabilidad** de este algoritmo es que se verifique

$$r = \frac{k}{h^2} \leq \frac{1}{2} \Longleftrightarrow k \leq \frac{1}{2}h^2 \tag{7.27}$$

Esta restricción hace que el método progrese lentamente. Por ejemplo, si $h = 0,01$ el máximo valor permitido de k es 5×10^{-5}. Si deseamos calcular una solución para $0 \leq t \leq 10$, entonces el número de pasos en el tiempo será 2×10^5 y el número de puntos de la malla debe ser superior a 20 millones, de modo que la elegancia y sencillez del método van acompañados de una cierta ineficacia, que se remedia en parte en los métodos implícitos que veremos más adelante. Ejemplos ilustrativos se pueden ver en (13). También se verifica el siguiente teorema de convergencia (puede verse en el libro de A. Iserles: A First Course in the Numerical Analysis of Diferrential Equaions. Cambridge Texts in Applied Mathematics, 1998).

Teorema 53 *Si* $0 < r \leq \frac{1}{2}$ *el método es convergente.*

Método implícito para la ecuación del calor. Utilizando diferencias regresivas para aproximar la derivada temporal de u, por Taylor se obtiene

$$u_t(x_i, t_j) = \frac{u(x_i, t_j) - u(x_i, t_{j-1})}{k} + \frac{k}{2} u_{tt}(x_i, t_j + \theta_j k) \; con \; -1 < \theta_j < 0 \tag{7.28}$$

y para la derivada segunda espacial la fórmula (7.20), con lo cual resultan las ecuaciones en diferencias

$$\boxed{\begin{array}{l} \frac{u_{i,j} - u_{i,j-1}}{k} - \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} = 0 \\ i = 1, 2, \dots, n-1; \; j = 1, 2, \dots, m-1 \end{array}} \tag{7.29}$$

estando ahora el error de truncamiento local dado por la expresión

$$\tau_{ij} = -\frac{k}{2} u_{tt}(x_i, t_j + \theta_j k) - \frac{h^2}{12} u_{xxxx}(x_i + \xi_i h, t_j) \tag{7.30}$$

Definiendo nuevamente la constante de Courant $r = \frac{k}{h^2}$, las ecuaciones (7.29) se expresan como sigue

$$\boxed{\begin{array}{l}(1+2r)u_{i,j} - ru_{i+1,j} - ru_{i-1,j} = u_{i,j-1} \\ (i = 1, 2, \ldots, n-1;\; j = 1, 2, \ldots, m-1)\end{array}} \tag{7.31}$$

puesto que $u_{i,0} = f(x_i)$ para $i = 1, 2, \ldots, n-1$, y $u_{n,j} = u_{0,j} = 0$ para $j = 1, 2, \ldots, m-1$, en forma matricial se tiene

$$\boxed{AU_j = U_{j-1} \Leftrightarrow U_j = A^{-1}U_{j-1} \; (j = 1, 2, \ldots, m-1)} \tag{7.32}$$

siendo

$$A = \begin{bmatrix} 1+2r & -r & \ldots & 0 & 0 \\ -r & 1+2r & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1+2r & -r \\ 0 & 0 & \ldots & -r & 1+2r \end{bmatrix}, \; U_0 = \begin{bmatrix} f(x_1) \\ \vdots \\ \vdots \\ \vdots \\ f(x_{n-1}) \end{bmatrix}$$

La matriz A es estrictamente diagonal dominante, tridiagonal simétrica y definida positiva, por lo cual puede utilizarse algún método específico para resolver sistemas lineales de este tipo (por ejemplo el método directo de Cholesky o los iterativos de Jacobi o Gauss-Seidel). Además, los autovalores de A son todos positivos y mayores que 1, por tanto existe A^{-1} y sus valores propios son todos positivos pero menores que 1, pues son los inversos de los de A, ahora un error e_0 en los datos iniciales se propaga al paso m en la forma $(A^{-1})^m e_0$ y el método resulta **estable** independientemente de la elección de r; por ello, a este tipo de métodos se le denomina **incondicionalmente estables**. Su orden es del tipo $O(k + h^2)$ suponiendo u de clase $C^{(4)}$ en un cierto dominio.

Para obtener un error de truncatura del tipo $O(k^2 + h^2)$ se utiliza el método de Crank-Nicolson que pasamos a describir y consiste en promediar las diferencias progresiva y regresiva.

Método de Crank-Nicolson para la ecuación del calor. Es un método de orden $O(h^2 + k^2)$ incondicionalmente estable; que se obtiene promediando la diferencia progresiva en el j-ésimo paso en t

$$\frac{u(x_i, t_{j+1}) - u(x_i, t_j)}{k} - \frac{u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)}{h^2} = 0$$

cuyo error local de truncamiento es

$$\tau_p = \frac{1}{2}ku_{tt}(x_i, t_j + \theta_j k) + O(h^2) \;\; (0 < \theta_j < 1)$$

y la diferencia regresiva en el (j+1)-ésimo paso en t

$$\frac{u(x_i,t_{j+1})-u(x_i,t_j)}{k}-\frac{u(x_{i+1},t_{j+1})-2u(x_i,t_{j+1})+u(x_{i-1},t_{j+1})}{h^2}=0$$

cuyo error local de truncamiento es

$$\tau_p=-\frac{1}{2}ku_{tt}(x_i,t_{j+1}+\epsilon_j k)+O(h^2)\quad(-1<\xi_j<0)$$

Como $0<\theta_j<1$ y $-1<\epsilon_j<0$, el esquema se construye suponiendo que $t_j+\theta_j k\cong t_{j+1}+\xi_j k$ resultando

$$\frac{u(x_i,t_{j+1})-u(x_i,t_j)}{k}-\frac{1}{2}\left[\frac{u(x_{i+1},t_j)-2u(x_i,t_j)+u(x_{i-1},t_j)}{h^2}+\frac{u(x_{i+1},t_{j+1})-2u(x_i,t_{j+1})+u(x_{i-1},t_{j+1})}{h^2}\right]=0$$

que en forma matricial, se expresa como

$$AU_{j+1}=BU_j,\quad j=0,1,2,\ldots,m-1$$

siendo

$$A=\begin{bmatrix}1+r & -\frac{r}{2} & \ldots & 0 & 0\\ -\frac{r}{2} & 1+r & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \ldots & 1+r & -\frac{r}{2}\\ 0 & 0 & \ldots & -\frac{r}{2} & 1+r\end{bmatrix}$$

y

$$B=\begin{bmatrix}1-r & \frac{r}{2} & \ldots & 0 & 0\\ \frac{r}{2} & 1-r & \ldots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \ldots & 1-r & \frac{r}{2}\\ 0 & 0 & \ldots & \frac{r}{2} & 1-r\end{bmatrix}$$

La matriz A es tridiagonal, simétrica y definida positiva, por lo tanto existe su inversa A^{-1}, y pueden emplearse los métodos especiales para resolver los sistemas lineales con este tipo de matrices de coeficientes.

Este método puede recordarse y representarse por medio de la denominada “molécula computacional” mediante la gráfica siguiente.

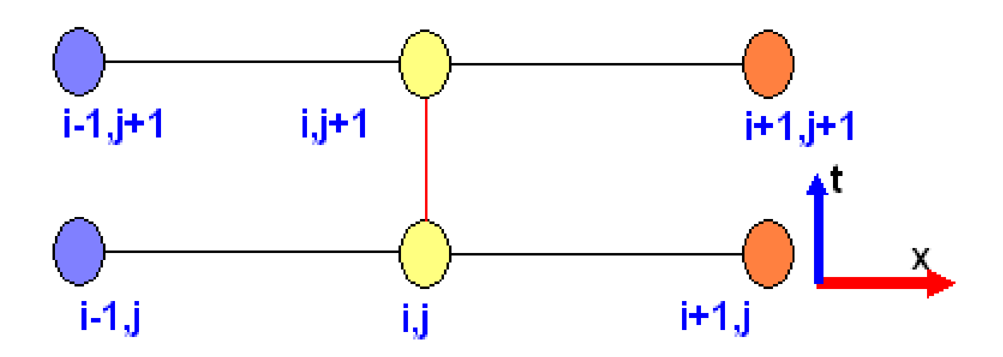


Figura 7.2: Molécula computacional del método de Crank-Nicolson

Observaciones: Los esquemas anteriores pueden aplicarse sin dificultad a las ecuaciones parabólicas generales. Así, dada la ecuación

$$u_t = a(x,t)u_{xx} + b(x,t)u_x + c(x,t)u + d(x,t)$$

con la condición inicial

$$u(x,0) = f(x), \quad (0 < x < 1)$$

y las condiciones en la frontera

$$u(0,t) = g_1(t), \; u(1,t) = g_2(t) \; (t \in \mathbb{R}^+)$$

Además, suponemos que verifican las condiciones de compatibilidad

$$f(0) = g_1(0), \; f(1) = g_2(0)$$

Utilizando las siguientes aproximaciones en diferencias

$$u_t(x_i,t_j) \cong \frac{u_{i,j+1} - u_{i,j}}{h}$$

$$u_x(x_i,t_j) \cong \frac{u_{i+1,j} - u_{i-1,j}}{h}$$

$$u_{xx}(x_i,t_j) \cong \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}$$

con la notación

$$a_{ij} = a(x_i,t_j), \; b_{ij} = b(x_i,t_j), \; c_{ij} = c(x_i,t_j), \; d_{ij} = d(x_i,t_j), \; r = \frac{k}{h^2}, \; s = \frac{k}{2h}$$

El esquema explícito resulta ser

$$u_{i,j+1} = (ra_{ij} - sb_{ij})u_{i-1,j} + (1 + kc_{ij} - 2ra_{ij})u_{ij} + (ra_{ij} + sb_{ij})u_{i+1,j} + kd_{ij} \quad (7.33)$$

y se tiene el siguiente resultado de estabilidad.

Proposición 9 *Si* $d = 0$ *y* a, b, c *son continuas con* $0 < A \leq a(x,t) \leq B$, $\mid b(x,t) \mid \leq C$, $-D \leq c(x,t) \leq 0$, $Ch \leq 2A$ *y* $k \leq \frac{h^2}{Dh^2+2B}$ *entonces el método (7.33) es estable.*

7.6. Ecuaciones hiperbólicas: la ecuación de ondas

Método explícito. Estudiemos la integración numérica de la ecuación hiperbólica unidimensional, dada por

$$u_{tt} = u_{xx} \ (0 < x < 1, \ t \in \mathbb{R}^+) \tag{7.34}$$

con las condiciones de frontera

$$u(0,t) = u(1,t) = 0 \ (t \in \mathbb{R}^+) \tag{7.35}$$

y las condiciones iniciales

$$u(x,0) = f(x), \ u_t(x,0) = g(x) \ (0 < x < 1) \tag{7.36}$$

Tomemos dos tamaños de paso constantes h, k, de modo que $1/h = n$ sea un número entero y k sea un número real positivo arbitrario. Así, tendremos definidos los nodos espaciales

$$x_i = ih, \ i = 0, 1, \ldots, n$$

y los tiempos discretos

$$t_j = jk, \ j = 1, 2, \ldots$$

La solución numérica vendrá dada por un conjunto de valores u_{ij}, que serán las aproximaciones de la solución exacta en los puntos: (x_i, t_j), ahora la fórmula de Taylor nos permite escribir (si la función $u(x,t)$ es de clase $C^{(4)}$ en su dominio)

$$\begin{array}{l} u_{tt}(x_i,t_j) = \frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{k^2} - u_{tttt}(x_i, t_j+\theta_j k)\frac{k^2}{12} \ (-1 < \theta_j < 1) \\ u_{xx}(x_i,t_j) = \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2} - u_{xxxx}(x_i+\xi_i h, t_j)\frac{h^2}{12} \ (-1 < \xi_i < 1) \end{array} \tag{7.37}$$

y sustituyendo en (7.34) las derivadas temporal y espacial por los primeros términos de estas aproximaciones, queda

$$\boxed{\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2}} \tag{7.38}$$

En tanto que el error de truncamiento está dado por la expresión

$$\tau_{ij} = \frac{1}{12}[k^2 u_{tttt}(x_i, t_j + \theta_j k) - h^2 u_{xxxx}(x_i + \xi_i h, t_j)] \tag{7.39}$$

Llamando $r = k/h$, el esquema (7.38) queda como sigue

$$\boxed{u_{i,j+1} = 2(1-r^2)u_{i,j} + r^2(u_{i+1,j} + u_{i-1,j}) - u_{i,j-1} \; (i = 1, \ldots, n-1; j = 1, 2, \ldots)} \tag{7.40}$$

Este esquema se puede representar por medio de la gráfica que sigue

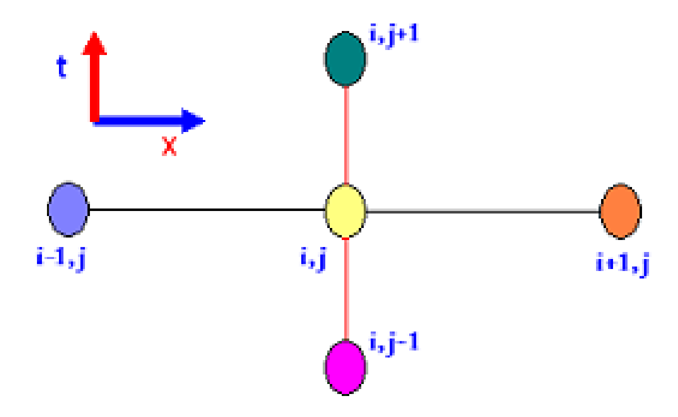


Figura 7.3: Molécula computacional del método explícito para la ecuación de ondas

Ahora, teniendo en cuenta las condiciones de frontera dadas en (7.35) y las condiciones iniciales dadas en (7.36), el método se puede expresar en la forma matricial

$$\boxed{U_{j+1} = AU_j - U_{j-1}} \tag{7.41}$$

siendo

$$A = \begin{bmatrix} 2(1-r^2) & r^2 & \ldots & 0 & 0 \\ r^2 & 2(1-r^2) & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 2(1-r^2) & r^2 \\ 0 & 0 & \ldots & r^2 & 2(1-r^2) \end{bmatrix}, \; U_j = \begin{bmatrix} u_{1j} \\ \vdots \\ \vdots \\ \vdots \\ u_{n-1,j} \end{bmatrix}$$

Para calcular U_{j+1} se requiere U_j y U_{j-1}, por tanto para iniciar el proceso (7.41) de cálculo de U_2, es necesario, además del valor de U_0 dado por las

condiciones iniciales, calcular U_1; aproximando $u_t(x,0)$ con el error de orden $O(k^2)$, similar al error de truncatura (7.39), lo que puede hacerse de la forma

$$\frac{u(x_i,t_1)-u(x_i,0)}{k} = u_t(x_i,0) + \frac{k}{2}u_{tt}(x_i,0) + \frac{k^2}{6}u_{ttt}(x_i,\theta k)\ (0<\theta<1)$$

Suponiendo que la ecuación de onda es válida en $t=0$, y que está definida $f''(x)$, se tiene

$$u_{tt}(x_i,0) = u_{xx}(x_i,0) = f''(x_i),\ \ (i=0,1,2,\ldots,n)$$

Además, si $f \in C^{(4)}([0,1])$, se verifica

$$f''(x_i) = \frac{f(x_{i+1})-2f(x_i)+f(x_{i-1})}{h^2} - \frac{h^2}{12}f^{(iv)}(x_i+\theta_i h),\ \ (-1<\theta_i<1)$$

y se tiene la aproximación

$$\frac{u(x_i,t_1)-u(x_i,0)}{k} = g(x_i) + \frac{k}{2h^2}(f(x_{i+1})-2f(x_i)+f(x_{i-1})) + O(k^2+kh^2)$$

es decir

$$u(x_i,t_1) = u(x_i,0) + kg(x_i) + \frac{r^2}{2}[f(x_{i+1})-2f(x_i)+f(x_{i-1})] + O(k^3+k^2h^2)$$

Luego, para representar los valores aproximados u_{i1} $(i=1,2,\ldots,n-1)$, se puede utilizar la ecuación en diferencias siguiente

$$u_{i1} = (1-r^2)f(x_i) + \frac{r^2}{2}[f(x_{i+1})+f(x_{i-1})] + kg(x_i) \tag{7.42}$$

Estabilidad del método. Por separación de variables discretas, se puede deducir que el método es **estable** si $r \leq 1$ (o equivalentemente si $k \leq h$). Puede construirse un método incondicionalmente estable como vemos a continuación.

Método implícito. Para obtener este método, basta con considerar las aproximaciones por diferencias finitas siguientes

$$\begin{aligned} u_{tt} &\cong \tfrac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{k^2} \\ u_{xx} &\cong \tfrac{1}{2}[\tfrac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{h^2} + \tfrac{u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1}}{h^2}] \end{aligned} \tag{7.43}$$

que conducen al esquema incondicionalmente estable

$$u_{i-1,j+1}-2(1+r^2)u_{i,j+1}+u_{i+1,j+1} = -u_{i-1,j-1}-u_{i+1,j-1}-4r^2u_{ij}+2(1+r^2)u_{i,j-1} \tag{7.44}$$

siendo ahora $r=\frac{h}{k}$, y cuya molécula computacional figura a continuación.

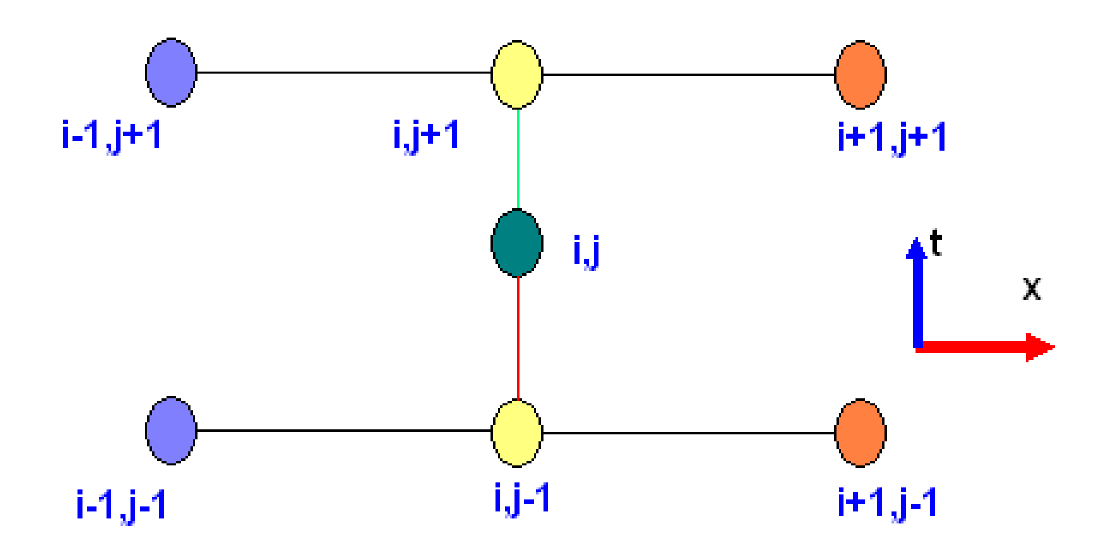


Figura 7.4: Molécula computacional del método implícito para la ecuación de ondas

7.7. Problemas resueltos

1. Resolver por un método en diferencias finitas centrales de segundo orden el problema de contorno:

$$y'' + y = t^2 + 2t + 3 \ (0 \leq t \leq 1); y(0) = 1, \ y(1) = 4$$

tomando $h = 0,25$, comparar la solución discreta obtenida con la exacta dada por $y(t) = (t+1)^2$, ¿a qué se debe la coincidencia de ambas soluciones?.

Solución. Como vimos en el capítulo 5, tomando $h = 0{,}25$ tenemos la malla de puntos $\{t_j = jh = 0{,}25j, j = 0, 1, 2, 3, 4\}$ del intervalo $[0, 1]$ y aproximaremos la derivada en los puntos t_j interiores a dicho intervalo $y''(t_j)$ como

$$y''(t_j) \cong \frac{y(t_{j+1}) - 2y(t_j) + y(t_{j-1})}{h^2}$$

cuyo error si se supone que la solución $y(t) \in \mathbb{C}^{(4)}([0, 1])$ puede escribirse en la forma

$$EDN(y''(t_j)) = -\frac{h^2}{12} y^{(4)}(\xi_j) \text{ con } \xi_j \in (t_{j-1}, t_{j+1})$$

Ahora, escribiendo estas aproximaciones para $j = 1, 2, 3$, y poniendo y_j en vez de $y(t_j)$ se tendrán las ecuaciones discretizadas que siguen

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} + y_j = t_j^2 + 2t_j + 3 \ \ (j = 1, 2, 3)$$

dado que $h = 1/4$ resultan ser

$$16y_{j-1} - 31y_j + 16y_{j+1} = t_j^2 + 2t_j + 3 \quad (j = 1, 2, 3)$$

y teniendo en cuenta los valores de frontera dados $y_0 = 1$ e $y_4 = 4$, que $tj = 0{,}25j$ y haciendo $j = 1, 2, 3$, las ecuaciones discretizadas pueden escribirse

$$\begin{array}{l} -31y_1 + 16y_2 = -12{,}4375 \\ 16y_1 - 31y_2 + 16y_3 = 4{,}25 \\ 16y_2 - 31y_3 = -58{,}9375 \end{array}$$

Resolviendo este sistema se obtienen $y_1 = 1{,}5625$, $y_2 = 2{,}25$, $y_3 = 3{,}0625$, que coinciden con los valores de la solución exacta dados por $y(0{,}25) = 1{,}25^2$, $y(0{,}5 = 1{,}5^2)$, $y(0{,}75) = 1{,}75^2$. Esta coincidencia se debe a que siendo la solución exacta $y(t) = (t + 1)^2$, un polinomio de segundo grado, su derivada cuarta es nula en todos los puntos, por tanto los errores de truncatura en las aproximaciones de las derivadas segundas utilizadas son nulos, en consecuencia si no cometemos errores de redondeo en la resolución del sistema lineal nos saldrá la solución exacta.

2. Resolver por un método en diferencias finitas centrales de segundo orden el problema de contorno:

$$2y'' - ty' + 2y = 6 - t \ (0 \le t \le 1); y(0) = 1, \ y(1) = 1$$

tomando $h = 0, 25$, comparar la solución discreta obtenida con la exacta dada por $y(t) = t^2 - t + 1$, ¿a qué se debe la coincidencia de ambas soluciones?.

Solución. Nuevamente dividimos el intervalo $[0, 1]$ en cuatro partes iguales, tomando $h = 1/4 = 0{,}25$ tenemos la malla de puntos de dicho intervalo $\{t_j = jh = 0{,}25j, j = 0, 1, 2, 3, 4\}$, y aproximaremos en los puntos interiores las derivadas segunda como en el ejercicio anterior y la primera como sigue

$$y'(t_j) \cong \frac{y(t_{j+1}) - y(t_{j-1})}{2h}$$

como vimos en el capítulo 5, si se supone que la solución $y(t) \in \mathbb{C}^{(3)}([0, 1])$ su error puede escribirse en la forma

$$EDN(y'(t_j)) = -\frac{h^2}{6} y^{(3)}(\xi'_j) \text{ con } \xi'_j \in (t_{j-1}, t_{j+1})$$

Llevando estas aproximaciones para $j = 1, 2, 3$ a la ecuación dada, y poniendo y_j en vez de $y(t_j)$ se tendrán las ecuaciones discretizadas que siguen

$$2\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} - t_j\frac{y_{j+1} - y_{j-1}}{2h} + 2y_j = 6 - t_j \quad (j = 1, 2, 3)$$

o también

$$(\frac{2}{h^2} + \frac{t_j}{2h})y_{j-1} + (-\frac{4}{h^2} + 2)y_j + (\frac{2}{h^2} - \frac{t_j}{2h})y_{j+1} = 6 - t_j \quad (j = 1, 2, 3)$$

dado que $h = 1/4$, teniendo en cuenta los valores de frontera dados $y_0 = 1$ e $y_4 = 1$, que $t_j = 0{,}25j$ y haciendo $j = 1, 2, 3$, las ecuaciones discretizadas resultan ser

$$\begin{aligned} &-62y_1 + 31{,}5y_2 = -26{,}75 \\ &33y_1 - 62y_2 + 31y_3 = 5{,}5 \\ &33{,}5y_2 - 62y_3 = -25{,}25 \end{aligned}$$

La soluciones de este sistema son $y_1 = 0{,}8125$, $y_2 = 0{,}75$ e $y_3 = 0{,}8125$ que coinciden con las soluciones exactas, dadas por $y(0{,}25) = 0{,}25^2 - 0{,}25+1 = 0{,}8125$, $y(0{,}5) = 0{,}5^2-0{,}5+1 = 0{,}75$ e $y(0{,}75) = 0{,}75^2-0{,}75+1 = 0{,}8125$. Nuevamente la coincidencia de las soluciones aproximada y exacta se debe a que se anulan los errores de truncatura o discretización y a que no cometemos errores de redondeo en los cálculos; los errores de truncatura se anulan por ser las derivadas tercera y cuarta que intervienen en las aproximaciones de las derivadas primera y segunda nulas, ya que la solución exacta en este caso es un polinomio de segundo grado.

3. Tomando $h = 0{,}5$, integrar el problema de contorno $y'' + \frac{1}{2}ty' - y = t$, $y(0) = y(2) = 2$ en el intervalo $[0, 2]$, puesto que la solución exacta de este problema está dada por $y(t) = (t - 1)^2 + 1$, ¿coincide la solución exacta con la numérica hallada?, ¿por qué?.

   Solución. En este trabajamos en el intervalo $[0, 2]$, que lo dividimos en cuatro partes iguales, tomando $h = 1/2 = 0{,}5$ tenemos la malla de puntos de dicho intervalo $\{t_j = jh = 0{,}5j, j = 0, 1, 2, 3, 4\}$, y aproximaremos en los puntos interiores las derivadas segunda y primera por aproximaciones de segundo orden, como se ha hecho en los ejercicios anteriores, y poniendo y_j en vez de $y(t_j)$ se tendrán las ecuaciones discretizadas que siguen

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} + \frac{t_j}{2}\frac{y_{j+1} - y_{j-1}}{2h} - y_j = t_j \quad (j = 1, 2, 3)$$

o también

$$(\frac{1}{h^2} - \frac{t_j}{4h})y_{j-1} - (\frac{2}{h^2} + 1)y_j + (\frac{1}{h^2} + \frac{t_j}{4h})y_{j+1} = t_j \quad (j = 1, 2, 3)$$

dado que $h = 1/2$, teniendo en cuenta los valores de frontera dados $y_0 = y_4 = 2$, que $t_j = 0{,}5j$ y haciendo $j = 1, 2, 3$, las ecuaciones discretizadas resultan ser

$$\begin{array}{l} -9y_1 + 4{,}25y_2 = -7 \\ 3{,}5y_1 - 9y_2 + 4{,}5y_3 = 1 \\ 3{,}25y_2 - 9y_3 = -8 \end{array}$$

La soluciones de este sistema son $y_1 = 1{,}25$, $y_2 = 1$ e $y_3 = 1{,}25$ que coinciden con las soluciones exactas, dadas por $y(0{,}5) = (0{,}5-1)^2+1 = 1{,}25$, $y(1) = (1-1)^2 + 1 = 1$ e $y(1{,}5) = (1{,}5-1)^2 + 1 = 1{,}25$. La coincidencia de las soluciones aproximada y exacta se debe a las mismas razones apuntadas en el ejercicio anterior.

4. Aproximar, por un método en diferencias finitas de segundo orden, tomando $h = k = 1$, la solución del problema de contorno:

$$u_{xx} + u_{yy} + (x+y)(u_x + u_y) - 2u = 0, \; \forall (x, y) \in R$$

$$u(x, y) = x^2 - y^2, \quad \forall (x, y) \in \partial R$$

siendo R el rectángulo $[0, 2] \times [0, 4] = \{(x, y) \mid 0 \leq x \leq 2, \; 0 \leq y \leq 4\}$ y ∂R su frontera. ¿Cuál puede ser la razón por la que coinciden la solución aproximada obtenida y la exacta?.

Solución. Sustituiremos las derivadas primeras y segundas por sus aproximaciones de segundo orden en los puntos interiores al rectángulo discreto $R = \{(x_i, y_j) \mid x_i = ih = i \; (i = 0, 1, 2), \; y_j = jk = j \; (j = 0, 1, 2, 3, 4)\}$, pues en este ejercicio hemos de tomar $h = k = 1$, y poniendo $u(x_i, y_j) = u_{ij}$, si la función solución $u(x, t)$ es de clase $C^{(4)}$ en su dominio, las fórmulas de Taylor nos permiten escribir

$$\begin{array}{l} u_{xx}(x_i, y_j) = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2} - u_{xxxx}(x_i + \xi_i h, y_j)\frac{h^2}{12} \text{ con } -1 < \xi_i < 1 \\ u_{yy}(x_i, y_j) = \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} - u_{yyyy}(x_i, y_j + \theta_j k)\frac{k^2}{12} \text{ con } -1 < \theta_j < 1 \\ u_x = \frac{u_{i+1,j} - u_{i-1,j}}{2h} - u_{xxx}(x_i + \xi_i' h, y_j)\frac{k^2}{6} \text{ con } -1 < \xi_i' < 1 \\ u_y = \frac{u_{i,j+1} - u_{i,j-1}}{2k} - u_{yyy}(x_i, y_j + \theta_j' k)\frac{k^2}{6} \text{ con } -1 < \theta_j' < 1 \end{array}$$

sustituyendo en la ecuación dada estas derivadas por los primeros términos de las aproximaciones anteriores para $i = 1$, $j \in \{1, 2, 3\}$, y teniendo en cuenta que $h = k = 1$ nos quedarán las ecuaciones discretizadas

$$u_{i+1,j} - 2u_{i,j} + u_{i-1,j} + u_{i,j+1} - 2u_{i,j} + u_{i,j-1} +$$

$$+(i+j)(\frac{u_{i+1,j}-u_{i-1,j}}{2}+\frac{u_{i,j+1}-u_{i,j-1}}{2})-2u_{ij}=0$$

o también

$$(1-\frac{i+j}{2})u_{i-1,j}-6u_{ij}+(1+\frac{i+j}{2})u_{i+1,j}+(1+\frac{i+j}{2})u_{i,j+1}+(1-\frac{i+j}{2})u_{i,j-1}=0$$

ahora, hacemos $i = 1$, $j = 1, 2, 3$ y tenemos en cuenta los valores en la frontera, con lo que obtenemos las ecuaciones

$$-6u_{11}+2u_{1,2}=-2u_{2,1}=-2\cdot(2^2-1^2)=-6$$

$$-\frac{1}{2}u_{1,1}-6u_{12}+\frac{5}{2}u_{1,3}=\frac{1}{2}u_{0,2}-\frac{5}{2}u_{2,2}=\frac{1}{2}(0^2-2^2)-\frac{5}{2}(0^2-0^2)=-2$$

$$-u_{1,2}-6u_{13}=u_{0,3}-3u_{2,3}-3u_{1,4}=-9-3\cdot(2^2-3^2)-3\cdot(1^2-4^2)=51$$

Resolviendo este sistema de tres ecuaciones lineales resultan $u_{11} = 0$, $u_{12} = -3$ y $u_{13} = -8$.

La coincidencia entre la solución aproximada obtenida y la exacta se debe a que en este caso la solución exacta del problema viene dada por la función $u(x, y) = x^2 - y^2$ (cómo fácilmente se puede comprobar), entonces a la vista de las expresiones anteriores de los errores de truncatura de las aproximaciones utilizadas paras las derivadas primeras y segundas, que incorporan derivadas terceras o cuartas de $u(x, y)$ en puntos intermedios, siendo estas nulas en cualquier punto debido a su expresión se anulan dichos errores, entonces si no cometemos errores de redondeo en los cálculos obtenemos la solución exacta.

5. Utilizando el método explícito para la ecuación del calor

$$u_t = u_{xx}\ (0 \leq x \leq 1, 0 < t)$$

con las condiciones de contorno

$$u(0, t) = u(1, t) = 0\ (0 < t)$$

y las condiciones iniciales

$$u(x, 0) = \sin \pi x,\ \ (0 \leq x \leq 1)$$

Obtener aproximadamente el valor de $u(x, t)$ para $t = 0{,}5$, tomando $h = 0{,}1$ y $k = 0{,}0005$. Asimismo, obtener dicha aproximación para $h = 0{,}1$ y $k = 0{,}01$. Comparar con la solución exacta dada por $u(x, t) = e^{-\pi^2 t} \sin \pi x$, ¿cuál es el valor de $r = \frac{k}{h^2}$ en cada caso?.

Solución. Como hemos visto en el apartado correspondiente a este método, el algoritmo para este problema viene dado por la fórmula

$$\begin{array}{l} u_{i,j+1} = u_{i,j}(1 - \frac{2k}{h^2}) + \frac{k}{h^2}(u_{i+1,j} + u_{i-1,j}) \\ (i = 1, \ldots, n-1;\ j = 0, 1, \ldots, m) \end{array}$$

que puede expresarse en forma matricial como sigue

$$U_j = AU_{j-1} = A^j U_0,\ j = 1, 2, \ldots, m$$

donde

$$A = \begin{bmatrix} 1-2r & r & \ldots & 0 & 0 \\ r & 1-2r & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1-2r & r \\ 0 & 0 & \ldots & r & 1-2r \end{bmatrix}$$

$$U_j = \begin{bmatrix} u_{1,j} \\ \vdots \\ \vdots \\ \vdots \\ u_{n-1,j} \end{bmatrix} \quad U_0 = \begin{bmatrix} \sin(\pi x_1) \\ \vdots \\ \vdots \\ \vdots \\ \sin(\pi x_{n-1}) \end{bmatrix}$$

en la que $r = k/h^2$, puesto que para $h = 0{,}1$ y $k = 0{,}0005$ es $r = 0{,}05 \leq 1/2$, el método es estable y convergente, resultando que para $t = 0{,}5$ hemos de dar 1000 pasos o sea se tendrá

$$U_{1000} = A^{1000} U_0$$

siendo A y U_0 las matrices 9×9 y 9×1 siguientes

$$A = \begin{bmatrix} 0{,}9 & 0{,}05 & \ldots & 0 & 0 \\ 0{,}05 & 0{,}9 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0{,}9 & 0{,}05 \\ 0 & 0 & \ldots & 0{,}05 & 0{,}9 \end{bmatrix} \quad U_0 = \begin{bmatrix} \sin(0{,}1\pi) \\ \vdots \\ \vdots \\ \vdots \\ \sin(0{,}9\pi) \end{bmatrix}$$

realizando los cálculos con wxMaxima resulta

$$U_{1000} = \begin{bmatrix} 0{,}0022865207865785 \\ 0{,}0043492209874396 \\ 0{,}0059861891352457 \\ 0{,}0070371873822617 \\ 0{,}0073993366973343 \\ 0{,}0070371873822617 \\ 0{,}0059861891352457 \\ 0{,}0043492209874396 \\ 0{,}0022865207865785 \end{bmatrix}$$

que podemos comparar con la solución exacta dada por

$$u(x_i, 0{,}5) = \begin{bmatrix} 0{,}0022224141785127 \\ 0{,}0042272829727624 \\ 0{,}0058183558564259 \\ 0{,}0068398875299933 \\ 0{,}0071918833558264 \\ 0{,}0068398875299933 \\ 0{,}0058183558564259 \\ 0{,}0042272829727624 \\ 0{,}0022224141785127 \end{bmatrix}$$

mostrando un error que verifica que $\| U_{1000} - u(x_i, 0{,}5) \|_\infty < 3 \cdot 10^{-4}$. Ahora, rehacemos los cálculos con $h = 0{,}1$ y $k = 0{,}01$, con lo cual resulta que $r = k/h^2 = 1 > 1/2$ el método no tiene por que converger, y realizando los cálculos se obtiene

$$A = \begin{bmatrix} -1 & 1 & \dots & 0 & 0 \\ 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 1 \\ 0 & 0 & \dots & 1 & -1 \end{bmatrix}$$

calculando nuevamente con wxMaxima, hemos de realizar 50 pasos para obtener la aproximación

$$U_{50} = A^{50} U_0$$

que resulta ser

$$U_{50} = \begin{bmatrix} 2490368{,}0 \\ 786432{,}0 \\ -524288{,}0 \\ 2621440{,}0 \\ 2621440{,}0 \\ 6291456{,}0 \\ 0{,}0 \\ 0{,}0 \\ 2228224{,}0 \end{bmatrix}$$

siendo ahora $\| U_{1000} - u(x_i, 0{,}5) \|_\infty = 6291455{,}993160113$, que muestra la no convergencia del método en este caso.

6. Consideremos la ecuación de ondas

$$u_{tt} = u_{xx} \ (0 \leq x \leq 1, 0 \leq t)$$

con las condiciones de contorno

$$u(0,t) = u(1,t) = t(1-t) \ (0 \leq t)$$

y las condiciones iniciales

$$u(x,0) = x(1-x) \ (0 \leq x \leq 1)$$

$$u_t(x,0) = 1 \ (0 < x < 1)$$

En $[0,1] \times [0, 0{,}5]$ considerar la malla (x_i, t_j) definida por $x_i = ih$, $t_j = jk$ siendo $h = \frac{1}{4}$, $k = \frac{1}{8}$; aproximar por un método explícito de segundo orden la solución en los puntos de dicha malla; comprobar que en este caso la solución numérica coincide con la analítica dada por $u(x,t) = x(1-x) + t(1-t)$.

Solución. Se consideran los nodos espaciales $x_i = i/4$ $(i = 0, 1, 2, 3, 4)$ y los tiempos discretos $t_j = jk = j/8$ $(j = 0, 1, 2, 3, 4)$, aproximando las derivadas $u_{tt}(x_i, t_j)$ y $u_{xx}(x_i, t_j)$ en los nodos interiores a dicha malla por aproximaciones de segundo orden se tendrá el algoritmo

$$u_{i,j+1} = 2(1-r^2)u_{i,j} + r^2(u_{i+1,j} + u_{i-1,j}) - u_{i,j-1}, \ (i = 1, 2, 3, \ j = 1, 2, 3) \ (*)$$

siendo ahora $r = k/h = 1/2 \leq 1$, por tanto se verifica la condición de estabilidad y teniendo en cuenta las condiciones de contorno e iniciales se tienen: $u_{0,0} = 0$, $u_{0,1} = 7/64$, $u_{0,2} = 3/16$, $u_{0,3} = 15/64$, $u_{0,4} = 1/4$, $u_{4,0} = 0$, $u_{4,1} = 7/64$, $u_{4,2} = 3/16$, $u_{4,3} = 15/64$, $u_{4,4} = 1/4$,

también necesitamos $U_0 = (u_{1,0}, u_{2,0}, u_{3,0})$ y $U_1 = (u_{1,1}, u_{2,1}, u_{3,1})$, el primero se calcula de las condiciones iniciales y está dado por $u_{1,0} = u(0{,}25, 0) = 0{,}25(1 - 0{,}25) = 0{,}1875$, $u_{2,0} = u(0{,}5, 0) = 0{,}5(1 - 0{,}5) = 0{,}25$ y $u_{3,0} = u(0{,}75, 0) = 0{,}75(1 - 0{,}755) = 0{,}1875$, para calcular el segundo suponemos que se cumple la ecuación de ondas sobre el eje OX, entonces, las componentes de U_1, pueden calcularse, ver la fórmula (7.42) de la teoría, en la forma

$$u_{i1} = (1 - r^2)f(x_i) + \frac{r^2}{2}[f(x_{i+1}) + f(x_{i-1})] + kg(x_i)$$

para $i = 1, 2, 3$, siendo ahora $f(x) = x(1 - x)$ y $g(x) = 1$, con lo cual se obtienen $u_{1,1} = 19/64 = 0{,}296875$, $u_{2,1} = 23/64 = 0{,}359375$ y $u_{3,1} = 19/64 = 0{,}296875$. Finalmente, aplicando el algoritmo (*), obtenemos

$$u_{1,2} = 3/8, \quad u_{2,2} = 7/16, \quad u_{3,2} = 3/8$$
$$u_{1,3} = 27/64, \quad u_{2,3} = 31/64, \quad u_{3,3} = 27/64$$
$$u_{1,4} = 7/16, \quad u_{2,4} = 1/2, \quad u_{3,4} = 7/16$$

que puede comprobarse coincide con la solución exacta dada en este caso por

$$u(0{,}25, 0{,}5) = 7/16,\ u(0{,}5, 0{,}5) = 1/2,\ u(0{,}75, 0{,}5) = 7/16$$

La razón de esta coincidencia está, nuevamente, en que los errores de redondeo son nulos por ser las derivadas cuartas respecto a x y t nulas y no cometer errores de redondeo.

7.8. Problemas y trabajos propuestos

Problemas propuestos:

1. Probar que los siguientes problemas de valor en la frontera tienen solución única

   a) $y''(t) = (5y(t) + sen3y(t))e^t \ (0 \leq t \leq 1), \quad y(0) = y(1) = 0$

   b) $y''(t) + e^{-ty(t)} + \sin(y'(t)) \ (1 \leq t \leq 2), \quad y(1) = y(2) = 0$

2. Utilizando el método en diferencias finitas centrales de segundo orden, resolver los siguientes problemas de valor en la frontera lineales

   a) $y'' + y' - y = t \ (0 \leq t \leq 1), \quad y(0) = -1, y(1) = -2$, tomando $h = 0{,}25$. Comprobar que la solución numérica coincide con la exacta y razonar el motivo de esta coincidencia.

b) $y'' + (t/2 - 3/2)y' - y = 0$ $(0 \leq t \leq 6)$, $y(0) = y(6) = 11$. Comparar el resultado numérico con la solución analítica, dada por la fórmula $y(t) = t^2 - 6t + 11$.

c) $y'' + 4y = \cos t$ $(0 \leq t \leq \pi/4)$, $y(0) = y(\pi/4) = 0$, tomando $h = \pi/12$.

3. Utilizar el método en diferencias centrales de segundo orden para aproximar la solución al problema de contorno

$$y' = -\frac{4}{t}y' + \frac{2}{t^2}y - \frac{2\log t}{t^2} \; (1 \leq t \leq 2), \quad y(1) = -\frac{1}{2}, y(2) = \log 2$$

tomar $h = 0{,}05$, tas obtener las ecuaciones discretizadas, utilizar Maxima para resolver el problema lineal correspondiente y comparar las soluciones obtenidas con los valores exactos.

4. Considerar el problema de Dirichlet sobre un rectángulo R, con vértices $(0,0), (3,0), (3,4), (0,4)$ y sea Ω el interior de R, se quiere resolver

$$\Delta u = 0 \;\; \forall (x,y) \in \Omega; u(x,y) = (x-1)^2 + (y-1)^2 \;\; \forall (x,y) \in R$$

mediante un método en diferencias finitas centrales, aproximar la solución en los puntos interiores de la malla determinada al tomar $h = k = 1$ (utilizar Maxima para resolver el sistema lineal).

5. Dada la ecuación de Poisson

$$u_{xx} + u_{yy} = xe^y \;\; (0 < x < 2, 0 < y < 1)$$

con las condiciones de frontera

$$u(0,y) = 0, u(2,y) = 2e^y \;\; \forall y \in [0,1]; u(x,0) = x, u(x,1) = e^x \;\; \forall x \in [0,2]$$

Aproximar la solución exacta (dada por $u(x,y) = xe^y$) en los puntos interiores de la malla determinada tomando $h = 1/3$ y $k = 1/5$ (utilizar Maxima para resolver el sistema lineal correspondiente).

6. Dado el problema para la ecuación del calor

$$u_t = u_{xx} \; (0 \leq x \leq 3, 0 < t)$$

con las condiciones de contorno

$$u(0,t) = 0, u(3,t) = 9 \; (0 < t)$$

y las condiciones iniciales

$$u(x,0) = x^2, \ (0 \leq x \leq 3)$$

Obtener aproximadamente el valor de la solución en los puntos de la malla obtenidos al tomar $h = 1$ y $k = 0{,}5$.

7. Utilizar el método explicito en diferencias centrales (para las derivadas espaciales) para aproximar la solución del problema

$$\begin{gathered} u_t = u_{xx} + (x-2)u_x - 3u \ \ (0 \leq x \leq 4, 0 < t) \\ u(x,0) = x^2 - 4x + 5 \ \forall x \in [0,4] \\ u(0,t) = u(4,t) = 5e^{-t} \ \forall t > 0 \end{gathered}$$

en la malla generada tomando $h = 1$ y $k = 0{,}1$.

8. Utilizar un método implícito para resolver el problema anterior tomando $h = 1$ y $k = 0{,}5$ (tomar sólo tres decimales al resolver los sistemas lineales obtenidos) y hallar el valor aproximado de $u(x_i, 1)$.

9. Sea el problema

$$\begin{gathered} u_{tt} = u_{xx} \ \ (0 \leq x \leq 1, 0 \leq t) \\ u(x,0) = x(1-x), \ u_t(x,0) = 1 \ \forall x \in [0,1] \\ u(0,t) = u(1,t) = t(1-t) \ \forall t \geq 0 \end{gathered}$$

en $[0,1] \times [0, 0{,}5]$ considerar la malla (x_i, t_j) definida por $x_i = ih$ y $t_j = jk$ siendo $h = 1/4$ y $k = 1/8$; aproximar la solución en los puntos de dicha malla (comprobar que en este caso los puntos de la solución numérica coinciden con los de la solución exacta dada por $u(x,y) = x(1-x) + t(1-t)$).

10. Aplicar un método implícito para obtener la misma aproximación del problema anterior.

Trabajos propuestos:
Se propone en este tema realizar alguno de los siguientes trabajos:

- Métodos de Galerkin y de Ritz para problemas independientes del tiempo.
- Introducción al método de los elementos finitos para ecuaciones diferenciales ordinarias.

- Introducción al método de los elementos finitos para ecuaciones en derivadas parciales elípticas.
- Introducción al método de los elementos finitos para ecuaciones en derivadas parciales parabólicas e hiperbólicas.

Bibliografía

[1] A. Aubanell, A. Benseny y A. Delshams: Útiles básicos de Cálculo Numérico. Labor, 1998.

[2] M. Calvo, J. I. Montijano y L. Rández: Curso de Análisis Numérico (Métodos de Runge- Kutta para la resolución numérica de ecuaciones diferenciales ordinarias) y Curso de Análisis Numérico (Métodos lineales multipaso para la resolución numérica de ecuaciones diferenciales ordinarias). Servicio de Publicaciones de la Universidad de Zaragoza, 1985.

[3] C. Conde Lázaro y G. Winter: MÉTODOS Y ALGORITMOS BÁSICOS DEL ÁLGEBRA NUMÉRICA. Reverté. Barcelona, 1990.

[4] J. R. Dormand: Numerical Methods for Differential Equations. CRC Press, 1996.

[5] J. D. Faires y R. L. Burden: Métodos numéricos (tercera edición). Thompson. Madrid, 2004.

[6] F. García Merayo y A. Nevot Luna: Análisis numérico (más de 300 ejercicios resueltos y comentados). Paraninfo. Madrid, 1992.

[7] M. Gasca: CÁLCULO NUMÉRICO I. UNED. Madrid, 1986.

[8] W. Gautschi: Numerical Analysis (An Introduction). Birkhauser. Boston, 1997.

[9] D. Greenspan y V. Casulli: Numerical Analysis for Mathematics, Science and Engineering. Addison-Wesley, 1988.

[10] D. Kincaid y W. Cheney: Análisis Numérico. Adisson-Wesley Iberoamericana. Delaware (E.E.U.U.), 1994.

[11] J. D. Lambert: Computational Methods in Ordinary Differential Equations. John Wiley and Sons, 1998.

[12] R. Théodor: Initiation a l' Analyse Numérique. Masson, Paris 1982.

[13] A. Tveito y R.Winther: Introduction to Partial Differential Equations (A computational Approach). Springer, 1998.

[14] A. Vigueras: Prácticas de Cálculo Numérico con Maxima, Universidad Politécnica de Cartagena, 2016.